

CONTRIBUTING AUTHORS

Robert C. Angell
A. Angus Campbell
Charles F. Cannell
Dorwin W. Cartwright
Glyde H. Coombs
Leon Festinger
Ronald Freedman
John R. P. French, Jr.
Roger W. Heyns
Robert L. Kahn
George Katona
Daniel Katz
Leslie Kish
Rensis Likert
Ronald Lippitt
Theodore M. Newcomb
Helen Peak
Keith Smith
Alvin F. Zander

RESEARCH METHODS IN THE BEHAVIORAL SCIENCES

MLSU - CENTRAL LIBRARY



17810CL

EDITED BY

LEON FESTINGER

Department of Psychology, University of Minnesota

DANIEL KATZ

Department of Psychology, University of Michigan

HOLT, RINEHART AND WINSTON

NEW YORK—CHICAGO—SAN FRANCISCO—
TORONTO—LONDON

17810

December, 1966

Copyright 1953 by Holt, Rinehart and Winston, Inc.

All rights reserved

2054153

Printed in the United States of America

13 14 15 16 17 18 19

Foreword

The discovery in our time that scientific methodology can be applied to human problems has revolutionized psychology and has seriously affected all branches of social science. This discovery, moreover, came during a period when problems of social adjustment had reached a critical point in the years of depression, of war, and of postwar crises. As a consequence, empirical and quantitative research in our field has seen unparalleled growth. This period of boom has naturally not been characterized by a high degree of order or of systematic development of theory and methodology. We have been too absorbed in doing research to plan thoroughly, to take stock of our progress, or to communicate our findings adequately and to inform one another of our techniques and approaches.

The first great break in this pattern came with the publication of *Studies in Social Psychology in World War II*. Stouffer and his collaborators took time out to set forth their findings and their methods in communicable form. These volumes were an excellent demonstration of the importance of the codification of research methods. In the early days of the social studies, there was justification for scholars to give the result of their insights and reflections without specifications concerning the ways in which they arrived at their interpretations, for in that period they were working more as intuitive artists than as scientists. But today, when we attempt

experimentation and quantification we have no excuse for failing to codify our procedures

One essential aspect of scientific technique is that it can be stated in a standard form and can be taught so that trained and competent investigators can apply it in the same fashion. We still have not achieved the degree of specification possible in the physical sciences. The method of the interview, for example, still combines art and science. It is only, however, through making our procedures explicit that we can test, criticize, and improve them.

Psychologists, social psychologists, and sociologists at the University of Michigan have felt fortunate in the favorable atmosphere for social research at their institution which has made possible many projects in the academic departments and the creation of the Institute for Social Research with its coordinate divisions of the Survey Research Center and the Research Center for Group Dynamics. Since this research development brought together many specialists, it seemed worthwhile to take advantage of their physical and psychological proximity to produce a book on methodology. Two purposes were dominant: (1) to help in the present trend toward codification of research techniques, and (2) to give graduate students in the field some understanding of the principles and procedures of modern methodology. The criterion for inclusion of methods was the degree of relevance to the problems of social psychology, and the criterion for exclusion was the availability of knowledge about a technique already standardized in another field. Thus, although factor analysis is a useful method in social science, the details of its application have already been described in statistical texts. Similarly, projective methods have been described in the personality context in which they are characteristically used. On the other hand, there has been a lack of detailed treatment of behavioral observation, of the quantitative analysis of qualitative materials, and of such major research settings as field studies and field experiments.

There has been another underlying purpose in the publication of this book. It is our belief that progress in any field must rest upon methods appropriate to that field. Although the basic logic of scientific methodology is the same in all fields, its specific techniques and approaches will vary depending upon the subject matter. In its early stages, social psychology was handicapped by a

lack of methods appropriate to its problems. In general, the recruits from the upper frontier of social science understood its larger problems but were unequipped as technicians to handle them. The technology came from the lower frontier of individual psychology, where there had been a long development in psychophysics, in laboratory methods, and in psychometrics.

The attempt to apply this type of technology to social psychology was much too literal and failed to consider the appropriateness of the technique to the problem under investigation. Hence the earlier efforts to test Freudian concepts were fruitless. In industrial psychology, precision measures of isolated motor performance were inadequate to cope with problems of fatigue and motivation. The item reliability technique of the psychometrician was no answer in itself to the need for measures of cognitive and motivational structure in dealing with social attitudes.

The real problem is not that techniques cannot be adapted to a variety of problems but that they tend to carry with them the type of thinking and even the concepts of the area in which they were developed. Thus, the experimental technique when first applied to social psychology attempted manipulation of the amount of social stimulation—*i.e.*, the sheer physical dimension, as in the alone and together experiments. The creation and manipulation of the specific social influences came as a later development. Thus, when old techniques are used in the social field they have to be adapted to the conceptual framework in which they are applied. Otherwise we shall find ourselves testing things other than the theories in which we are interested. Moreover the special problems of our field call for new approaches and new techniques. The traditional measurement procedures involved assumptions not necessarily met by social data. The development of new scaling methods and of nonparametric statistics are hopeful signs of progress in this respect.

Finally, the social researcher should consider his research design from the point of view of testing the significant theories in his own field rather than from the frame of reference of what he would be doing if he were determining a sensory threshold. It is our conviction that methodologies need to be written for the field of social psychology itself.

Most of the contributors to this volume are social psychologists

This means that the problems they discuss tend to be taken from the field of social psychology. It is our belief, however, that there are many areas in the social sciences to which the approaches and methods described in the following pages will have application. For areas in the social sciences which deal with relationships between group indices without reference to intervening variables, these methods may need the same sort of adaptation to meet the criterion for appropriateness as was demanded in the field of social psychology when it took over techniques from individual psychology.

A cooperative undertaking of this sort requires not only the assistance of the contributors of the chapters which follow but the support of their colleagues. We are indebted to Donald Marquis, who participated in the planning of the project and who bears much of the responsibility for the circumstances which made the book possible. Other participants in the project not formally represented in the following chapters were Eugene Jacobson, Lowell Kelly, Charles Metzner, Ian Ross, and Guy Swanson. In most books there is generally one person who carries the brunt of editorial work, and it has been our good fortune to have had Mrs. Emily Willerman for this role, which she has carried out with unusual devotion and competence.

University of Michigan
June 21, 1953

L. F.
D. K.

Contents

INTRODUCTION

- The Interdependence of Social-Psychological
Theory and Methods: A Brief Overview 1

Theodore M Newcomb

PART I

RESEARCH SETTINGS 13

1. The Sample Survey: A Technique for
Social Science Research 15

A Angus Campbell and George Katona

2. Field Studies 56

Daniel Katz

- 3. Experiments in Field Settings** 98
John R. P. French, Jr

- 4. Laboratory Experiments** 136
Leon Festinger

PART II

PROCEDURES FOR SAMPLING 173

- 5. Selection of the Sample** 175
Leslie Ktish

PART III

METHODS OF DATA COLLECTION 241

- 6. Problems of Objective Observation** 243
Helen Peak

- 7. The Use of Documents, Records, Census
Materials, and Indices** 300
Robert C. Angell and Ronald Freedman

- 8. The Collection of Data by Interviewing** 327
Charles F. Cannell and Robert L. Kahn

9. Observation of Group Behavior	381
<i>Roger W. Heyns and Alvin F. Zander</i>	

PART IV

THE ANALYSIS OF DATA	419
----------------------	-----

10. Analysis of Qualitative Material	421
<i>Dorwin P. Cartwright</i>	

11. Theory and Methods of Social Measurement	471
<i>Glyde H. Coombs</i>	

12. Distribution-free Statistical Methods and the Concept of Power Efficiency	536
<i>Keith Smith</i>	

PART V

THE APPLICATION OF RESEARCH FINDINGS	579
---	-----

13. The Utilization of Social Science	581
<i>Rensis Likert and Ronald Lippitt</i>	

INDEX	647
-------	-----

The Interdependence of Social-Psychological Theory and Methods: A Brief Overview

Theodore M Newcomb

It is a truism that no research results are any better than the methods by which they are obtained. Behind the platitude, however, lie many complexities. Between the initial sensing of a problem and the final application of research results to that problem, there lies many a choice, as the reader of this volume will discover. At each dividing of the path, moreover, there are diverse criteria for deciding what is "better." Just as Moliere's M Jourdain was astonished on discovering that he had been speaking prose all his life, so not a few experienced researchers in social psychology will be amazed, on completing the baker's dozen of chapters that follow, to learn how many decisions they have been making these many years—with or without knowing it. It is one of the objectives of this volume to create a more general awareness of the existence of the choice points and of the criteria by which decisions may be made.

No article of faith in the scientist's credo is more elementary than his empiricist conviction that, if he learns how to ask the proper questions of "nature," he can formulate the principles accord

ing to which "nature" behaves. If our questions are not properly put—i.e., if our observations are not suitably made—in the first place, no amount of interpretative ingenuity at a later stage will enable us to reach our research objectives. Such methodological problems as devising interview schedules, selecting a sample of persons or of written words, manipulating a variable in the laboratory, or constructing an objective test are all problems of ensuring that the questions which we put to "nature" will be maximally suitable to elicit the answers required by our objectives. Problems of scaling, categorizing, discovering covariation, and testing the significance of differences are not merely matters of "translating" data already obtained, they are basic in the sense of determining—whether we know it or not—the kinds of questions which we are putting. What ever truth or falsity inheres in our research findings is quite as much a function of the questions we have elected to ask through our selection of methods as of the logic we have applied to the data elicited by our questions.

The kinds of research problems described in the following chapters have been attacked within a very limited time space setting. The methodological weapons which have been devised have, like military weapons at a given time and place in history, been conditioned by their setting. Some aspects of this setting impinge alike upon every variety of research into human behavior, some have had a special impact upon social research, and still others have influenced in specific ways the somewhat dimly demarcated field of social psychological research. This brief introductory chapter attempts to point to some of the contemporary methodological problems for social research in general and for social psychology in particular, and to note the position of the social psychologist in the confraternity of social researchers, as he borrows from and lends to his fellow members in the common enterprise.

SOME COMMON PROBLEMS IN SOCIAL RESEARCH

Social scientists face certain human problems which the natural scientist is spared. As we shall note particularly in Part I these

problems begin with getting access to persons as sources of data. It has too often gone unrecognized that the problem is not just one of avoiding refusals, whether by doorstep respondents, by student subjects, or by representatives of organizations. Such motivational factors on the part of the interviewees as those noted by Charles F. Cannell and Robert L. Kahn in Chapter 8 must be taken into account not only by interviewers but also by analysts in interpreting responses. In a very literal sense, moreover, the conditions under which clients and respondents first agree to participate in an investigation determine the nature of the eventual findings. Both respondent and investigator, whether they know it or not, are taking roles. The respondent's initial structuring of this role relationship will influence, in conscious and in unconscious ways, both the fullness and the content of his later responses—just as initial orientations toward any object influence later behavior in relation to it. As Ronald Lippitt and Rensis Likert note in Chapter 13, the process of research planning must include these facts of social life. The investigator has made a research decision, whether he knows it or not, when he first approaches a client, a subject, or a respondent, and he sets the stage for later decisions as he continues or modifies the initial role relationship.

One aspect of the role relationship between investigator and subject is an ethical one. In Chapters 3 and 4, some of the uses of temporary dissembling are noted, together with the responsibilities imposed upon the investigator who uses them—responsibilities to colleagues, who may later have to pay a heavy price for his laxity, and to the 'consumers' of his research findings, as well as to the subjects or respondents most directly involved. The last of these obligations is probably most easily met, as Chapter 4 suggests, most subjects accept without resentment the fact of having been duped, once they understand the necessity for it. Nevertheless, frequent use of what is regarded as deceit may lead to community shared expectations which undermine the necessary relationship of confidence between investigator and subject. The attitudes of subjects recruited from such a community may be such as to influence their responses in ways quite unsuspected by the investigator and thus, perhaps, invalidate his findings or his interpretations.

On a fundamental level, every social researcher, whether or not

he resorts to methods of dissembling stands in an ethical relationship to the community in which he works. As Doctor A. T. M. Wilson, of the Tavistock Institute, has said, It isn't so much that honesty is the best policy—it's the only possible policy. The honest investigator, who knows (or should know) more than the client about the consequences for the client of participation in the research, must not only not take advantage of his wider knowledge but must actually seek to turn it to the client's advantage. The temporary duping of laboratory subjects does not necessarily violate this concept of honesty, whereas sheer thoughtlessness on the part of an investigator who would never think of lying to a client may violate it fundamentally.

Subjects and clients, as well as investigators, have personal values which are apt to become involved in the research process. To assume that these are freely exploitable is, to quote Dr. Wilson again, as unreasonable as for a surgeon to accost a healthy man with the request: Pardon me, sir, may I rip open your abdomen in the interests of science? The fundamental ethical principle, in short, is based upon the recognition that in the long run the achievement of the investigator's objectives is dependent upon his respect for the client's values. The principle, thus stated, is just as applicable to the natural as to the social sciences, but its impact upon the latter is much more direct.

There is one of the occupational hazards of social research which can probably be avoided eventually, though it seems almost inevitable in the early stages of research methodology. This danger is best labeled, perversely enough, the temptation to anthropomorphize about humans. In its most extreme form—now, happily outgrown by most of us—it results in observations obtained by sheer intuition or empathy. In its more common contemporary form—and perhaps its most dangerous form since it seems natural and indeed inevitable—it results not in observations but in concepts and variables selected by reason of the fact that the investigator as well as the object of investigation is a human being. As a human being, as a previous participator in situations like those which he wishes to observe, he almost inevitably conceptualizes in anthropomorphic manner. The tendency is perhaps forgivable—relatively so at least—in the selection of our dependent variables, after all, our dependent variables are fairly close to the problems which set

our research mechanisms in action. But phenotypic phenomena are not necessarily the most significant ones to observe, even as dependent variables; and as intervening or independent variables they have, it seems to this writer, little better than chance likelihood of being the most significant ones. In clinical psychology, for example, a Rorschach W may be a more significant variable than the more humanly phrased "social expansiveness." Just so, in social science we probably have more to gain by identifying genotypic X's (with or without human-sounding labels) than by seeking to refine our measures of readily observable, human phenotypes. Helen Peak, in Chapter 6, has examined some of the properties of significant variables in terms of "functional unity."

Another problem of which social scientists of nearly every stripe are becoming increasingly aware has to do with the decisions they make when they employ a given process of measurement. In Chapter 5, Leslie Kish points to some of the consequences of using one device rather than another at various stages of sampling procedures. Keith Smith, in Chapter 12, points out some of the assumptions involved in the use of what may be our favorite statistical procedures and suggests alternatives which many of us will find more appropriate, once we are aware of the nature of the statistical decisions we have been making.

Clyde H. Coombs, in Chapter 11, goes to the very roots of the question "What is the nature of measurement itself?" Since we are necessarily doing something when we transform "real events" into numbers—whether at the stage of making observations or at the stage of analysis—it behooves us to know what we are doing. The requirements of this transformation process, together with certain properties of the events which the social scientist studies, confront him with a special dilemma. This chapter is characteristic of the tone of the entire book: instead of presenting recipes to be followed, it seeks to understand the logic of a type of problem which social scientists frequently meet. Since for the social scientist every investigation situation includes a large component of uniqueness, and since (as Dorwin Cartwright notes in Chapter 10, for example) the decisions made at every step of the investigation process are dependent upon decisions made at other steps, the investigator himself must construct his own blueprint.

SPECIAL PROBLEMS IN SOCIAL PSYCHOLOGICAL RESEARCH

In addition to the hazards which beset all kinds of social research, social psychology is subject to some 'special disabilities' *sui generis*. Many of these stem from its interdisciplinary parentage and its still ambiguous boundaries. From the point of view of range and richness social psychology has gained much by its generous borrowings from individual psychology, sociology, cultural anthropology, and psychiatry. But the cost has been heavy. We draw upon a wide and only loosely integrated variety of concepts. Our sources of data are as diverse as the analytic couch, the laboratory, the playground, the factory, the community, and the random sample of adults in a total society. It is likely, indeed, that many social psychologists are not even aware that they have at their disposal so broad a range of materials as those noted by Robert C. Angell and Ronald Freedman in Chapter 7. It is not surprising, therefore, that we have recourse to a wide range of methods in making observations, in isolating, measuring, and controlling our variables, and in analyzing our data. Under such conditions the difficulties of discriminating among settings and methods of apparently equal relevance and serviceability increase exponentially with the range of alternatives. Social psychological research seems, at midcentury, to be peculiarly subject to these difficulties.

There is, of course, nothing intrinsically undesirable in having recourse to a wide range of settings and methods, diversity, like adversity, has many uses. Several of the following chapters note possible ways—or even necessary ones—of taking advantage of this situation. Thus it is pointed out in one or more of the chapters in Part I that field studies have called attention to an increasingly wide range of variables which are manipulable or controllable in the laboratory, that field studies and surveys can supplement each other in significant ways, that certain kinds of problems are furthered by a planned sequence of quite different methods, and that, finally, genuinely programmatic research will very often necessitate such planned sequences. In somewhat similar vein Leon Festinger, in Chapter 4, discusses the special problem of 'artificiality' in the laboratory as contrasted with the realness of other situations.

It may happen, of course, that a specific *investigator* will specialize in a given method sheerly because of the conveniences or the necessities of division of labor. But this is not the same thing as specialization of method for a given *problem*. Hunches emerging from one sort of methodological study will often need to be tested by other methods, and results confidently confirmed by one method may turn out to be so loaded with the situational factors necessarily associated with the use of that method as not to be confirmable at all when other method-situation complexes are used. The more "basic" the problem, in general, the more essential it is that it be investigated programmatically by a designedly wide range of methods.

Another, and doubtless closely related, handicap under which social psychology currently suffers is that of too rapid growth. If its development has not been exactly hypertrophic, the tendency has certainly been to incorporate more than has been digested. John R. P. French, Jr., observes in Chapter 3, for example, the relatively low levels of abstraction at which social-psychological propositions are of necessity frequently formulated. And Daniel Katz notes, in Chapter 2, the relative rarity with which social-psychological investigations have been replicated. To an unknown extent both of these conditions are attributable to considerations of methodology. When we shall have achieved more satisfactory and more "standardized" methods of investigation, with clearer criteria concerning the range of appropriateness of various methods and tools, we shall, of course, be in a *much better position to make genuinely comparable studies*, which are a necessary precondition to high-level abstractions. And, similarly, a more satisfactory armamentarium of tools and methods will facilitate the replication of significant studies.

One's assessment of the consequences of these and other present shortcomings of social psychology will depend upon the nature of one's hopes and expectations and upon one's definition of the field. If one regards social psychology as an applied field, one will be dissatisfied with the present state of affairs primarily on grounds that the applicability of a given principle to a given situation is highly uncertain. But if one looks at social psychology as itself constituting a body of theory, the nature of one's concern will be quite different. Furthermore, the sources to which one looks for the improvement of methods will vary considerably, according as one's

definition of the field stresses its applied or its theoretical aspects

This brings us to the reciprocal of our initial truism. No method is any better than the theory by which it is tested. This point, touched upon in several chapters, is most explicitly made by Helen Peak, in Chapter 6, and Roger W. Heyns and Alvin F. Zander in Chapter 9. To the extent that our objectives center about applicability, we shall be content with empirical tests of their adequacy. This way, however, does not lead in the direction of generalizability and high levels of abstraction. Particularly in a field characterized by diverse kinds of settings and multiple divergent situational determinants, this way leads, at best, to a cookbook-like compendium of directions. Even from the point of view of applicability it is a discouraging way, since the possible concatenations of situational variables are almost infinite.

Conversely, to the extent that we are moving toward objectives of high level generalizability, we shall look to theoretical tests of our methods as the crucial ones—*i.e.*, crucial in ways such as those suggested by Helen Peak. But since this book is devoted to social psychological methods, the question is thus raised concerning the existence or even the possibility of a genuinely social psychological theory.

To such a question there would probably be no unanimous reply from the several authors of this volume—much less from the somewhat miscellaneous body of their colleagues who refer to themselves as social psychologists. But, among the present authors at least, the differences spring from nothing more consequential than differences in the use of labels and in notions of what are proper boundary lines among disciplines. All would agree that the processes by which persons relate themselves to one another and simultaneously to other aspects of their environment occur in orderly fashion and are subject to scientific investigation. This writer likes to apply the label *social psychology* to this area of investigation. He may differ with some among his colleagues as to the degree to which the principles accounting for these orderly processes represent applications of principles borrowed from neighboring disciplines. If so, there are differences among us as to the independent status of social psychology but not as to the importance of establishing the principles at the highest possible level of abstraction.

SOCIAL PSYCHOLOGICAL AS RELATED TO OTHER RESEARCH METHODS

Suppose it be agreed that there exists an area of theoretical importance which does not lie entirely within the confines of either sociology or psychology. Let it be accorded a quasi-independent status and labeled, for the time being, 'social psychology'. And suppose it be also agreed that there exists a body of methods characteristically used by social psychologists, not all of which are used by either sociologists or psychologists. Does it follow that there must also exist a quasi-independent body of social psychological methods? It might follow from the demands of an exaggerated professional pride but not from those of logic. Such distinguishability as social psychology has from its neighboring disciplines inheres in whatever distinctiveness characterizes the problems to which it addresses itself, the methods by which it attacks them might or might not be distinctive. In actual practice, there is little distinctiveness of method.

It is evident, even from a quick scanning of the Table of Contents, that the methodological problems treated in this book are by no means the exclusive or necessarily the primary concern of social psychologists. If there are basic methodological problems in the planning of laboratory experiments, for example, social psychologists have not been the first nor will they be the last to be faced by them. Social psychologists, being relative latecomers on the scene of research in human behavior, have been pioneers in few of the areas here represented, and in some instances their contributions have been relatively minor ones. Strictly speaking, there are probably no social psychological methods as such. The opening chapter, by Angus Campbell and George Katona, for example, emphasizes the interdisciplinary nature of survey methods.

Nevertheless, the total contribution of social psychologists to social research methodology has been considerable. Precisely because their methodological problems have not been entirely distinctive, they have been in a position to lend. What they have had to lend has not been of their sole creation. They have had to borrow, but they have also had to adapt what they have borrowed, and in adapting they have invented. It is their marginal status which has made

this both necessary and possible. Having one foot in pioneer territory, they felt the necessity to invent. Having the other foot in territory at least partly explored, they had some awareness of methods available to be borrowed and adapted.

Nowhere is this better illustrated than in the still unfinished problem of quantifying qualitative data. Sociologists and psychologists had long since developed their indices of discrete, objectively countable events and—borrowing from biologists—their statistical devices for analyzing their data. But many of the data crucial for the social psychologist's purposes were not discrete and countable. Being unwilling to give up the advantages of quantitative control, and at the same time aware of the unsuitability of standard parametric statistics, the social psychologist was forced to reconsider the whole research process. This meant keeping in mind from first steps to last the interdependence of data-gathering methods and the statistical methods by which they were to be analyzed. Awareness of such problems has resulted not only in the complete recasting of the methodology of attitude research but also in the development of new theories of scaling and even of measurement itself.

The marginality of the social psychologist's concerns may force him to become methodologically inventive, but it is no guarantee of the adequacy of his inventions. There must be selectivity as well as inventiveness, and selectivity presupposes a keen sense of the appropriateness of methods to problems. But criteria of appropriateness may conflict with other criteria. For example, most social psychologists have been fairly well schooled in the axioms of rigorous objectivity. If we cannot measure something we are tempted, as Clyde H. Coombs points out in Chapter 11, to go ahead and measure it anyhow. Such methodological compulsiveness defeats our basic objective of remaining maximally faithful to the events which we observe. The proper solution to this dilemma of rigor *vs* faithfulness lies not in abandoning either objective but in reassessing the means by which rigor is attainable, given that a certain sort of event is to be investigated. True rigor lies not in slavishly borrowing the standards developed for other purposes but in combining maximum faithfulness to empirical events with maximum reproducibility of procedures. Specific standards of rigor are relative, and not transferable.

If social psychology has led to methodological inventiveness

it has not been because of any jurisdictional label which its practitioners have worn. Rather, sensitivity to characteristically social-psychological problems has mothered invention because it presupposes the kind of cross-disciplinary familiarity which fosters discrimination among different sorts of data and among different sorts of methods. This, in turn, enhances the sense of appropriateness of methods to data and of both to basic theoretical problems.

SIGNPOSTS TO METHODOLOGICAL DEVELOPMENT

If it is true that research results are no better than the methods by which they are obtained, it is also true that the maturity of a discipline can be gauged by the spread of methodological sophistication among its practitioners. By this test, social-psychological research is perhaps not very far developed, but its rate of development seems nonetheless encouraging. The publication of three works within as many years—the present one, the 1951 volumes entitled *Research Methods in Social Relations* (1), and a forthcoming *Handbook in Social Psychology* (3), each devoted in whole or in part to methods of social-psychological research—is both a sign of a felt need for such sophistication and a probable cause of its spread. One of the necessary conditions for the growth of science is the replicability of its findings; this depends upon standardization and codification of its methods, which, in turn, depend upon communication through books such as this one.

Viewed in developmental perspective, the present position of social-psychological research thus seems a reasonably promising one. Among other factors which will determine its position in years to come, one is of particular importance: methodological research. At more than one point in this book the contributor has been forced to rely upon lore—i.e., common sense and the cumulative “wisdom” of experienced researchers. This, too, should be communicated, but it is no substitute for controlled research procedures. The “wisdom,” moreover, may turn out to be very unwise indeed. There has been considerable difference of opinion, for example, as to the advantages sometimes claimed for the face-to-face interview. But the necessity for basing our procedures upon mere opinions begins to disappear

when it is shown that certain specific differences, as predicted by a theory that barriers to communication vary with role relationships, distinguish interview responses and responses to printed questionnaires by the same persons (2).

The authors of this volume have been able to draw upon a limited pool of methodological research findings, to which they have themselves contributed. As social psychologists apply these research procedures to their own methodological problems, this pool will continue to expand.

BIBLIOGRAPHY

1. Jahoda, M , Deutsch, M , and Cook, S W., *Research methods in social relations* New York: Dryden Press, 1951.
2. Kahn, R L *A comparison of two methods of collecting data for social research: The fixed alternative questionnaire and the open ended interview* Ph D thesis, Univ of Michigan, 1952
3. Lindzey, G (ed) *Handbook of social psychology* To be published by Addison Wesley

PART I

Research Settings

Empirical research in social psychology, sociology, and related areas proceeds in a variety of settings and contexts. The choice of setting for any research project is generally guided by the nature of the questions being asked and the degree of control desired.

The variety of settings with which we must, then, concern ourselves may be ranged along a continuum from broad to narrow. We have chosen, for the sake of convenience, to divide this range arbitrarily into four sections. The boundaries between any two adjacent sections are never perfectly clear and sharp, but the division seems to have some validity in that different considerations and different techniques become important as we move from one to the other.

The broadest setting is, of course, one in which a large and perhaps spread out population is designated for investigation. Here survey methods are usually employed. Intensive study of field situations usually requires a narrower setting, such as a certain community or organization or industrial plant. When experiments are to be done in real life settings, it is usually necessary to narrow the

The Sample Survey: A Technique for Social- Science Research

A Angus Campbell and George Katona

Many research problems require the systematic collection of data from populations or samples of population through the use of personal interviews or other data gathering devices. These studies are usually called surveys, especially when they are concerned with large or widely dispersed groups of people. When they deal with only a fraction of a total population (or universe) a fraction representative of the total, they are called sample surveys.

The basic survey procedure, as used in the social sciences, is made up of a combination of techniques which have been developed in various research disciplines. The procedures of interviewing, for example, are based largely on the experience of psychologists, anthropologists, and others who used the personal interview both as a research tool and as a means of diagnosis or therapy long before it was adapted for survey use. Techniques of scaling and other methods of measurement have been borrowed from both sociology and psychology. Sampling methods have come in part from agricultural economics. Methods of content analysis have been drawn from

a variety of fields, including political science. Techniques of statistical analysis of mass data are common to all fields of quantitative research in the social sciences.

The survey instrument is not the specific method of any one social science discipline, and it is broadly applicable to problems in many fields. It is this capacity for wide application and broad coverage which gives the survey technique its great usefulness in the behavioral sciences.

Surveys depend on direct contact with those persons, or a sample of those persons, whose characteristics, behaviors, or attitudes are relevant for a specific investigation. Thus, the survey method differs from research carried out in libraries or archives by studying, re-grouping, and analyzing records compiled for other purposes.

The survey technique is used only when the desired information cannot be obtained more easily and less expensively from other sources. It would be very inefficient, for example, to conduct a sample survey to determine the number of passenger cars in use in the United States. Since American automobile owners must register their cars every year, a study of the files of the state licensing bureaus can provide such information much more rapidly and reliably. But no information is available from any records on the occupations, intentions, habits, or other characteristics of automobile owners. Similarly, annual data are compiled, mostly from records kept for practical and legal purposes, about the total national income or the national birth rate. Information about the income or birth rate of such subgroups of the population as inhabitants of large cities, skilled workers, or college graduates can be computed for certain years from the Decennial Census (which represents, in a sense, an application of the survey technique). But sample surveys are necessary to find the answers to such questions as: How many and what kind of American families had an income of less than \$1000 in each of several successive years? or: Are people with high incomes more optimistic about the future than people with low incomes? or: What are the differences in the birth rate of people of different religious affiliations?

Sample surveys can be undertaken only if the people selected as respondents are able and willing to give the desired information. It would not be very rewarding to ask a sample of the population to report their basic metabolic rate or to tell the interviewer whether they have repressed hostility feelings toward their father. Such

information requires specialized techniques of determination, and few people would be able to report these facts with any degree of validity. Furthermore, it cannot always be assumed that people will be willing to give information which they may reasonably be expected to have. Any information which might embarrass or incriminate the respondent is likely to be suppressed or distorted. Nevertheless, the tolerance of respondents for inquiries into their personal affairs is surprisingly high provided the interview is conducted skillfully and tactfully. This willingness of most respondents to give detailed information about themselves cannot be taken to mean, however, that the people who are asked questions in a survey will give answers with the impersonality of an electric tabulator. It is often easier to specify the data which a survey hopes to assemble than it is to devise and carry out the interview through which the data can be successfully obtained.

Surveys vary greatly in their scope, their design, and their content. As in any other research, the specific characteristics of any survey will be determined by its basic objectives. The statement of the essential questions which the research is intended to investigate delineates in large part the universe to be studied, the size and nature of the sample, the type of interviewing to be used, the content of the questionnaire, the character of the coding, and the nature of the analysis. Specific survey methods vary according to specific survey objectives.

TYPES OF UNIVERSES SURVEYED

The most widely known sample survey conducted in this country is the poll conducted by the Gallup organization. This journalistic enterprise has been asking questions of the American public since 1936, and its reports are a familiar weekly feature of many newspapers. Usually the Gallup Poll is intended to represent the total adult population; in its pre-election straw votes it attempts to represent the population of eligible voters.

Although the Gallup Poll has been the most highly publicized survey of the national population numerous other nation-wide surveys are conducted each year. Perhaps the most ambitious of these is the Current Population Survey, carried out by the Bureau of the

Although geographically defined populations of these kinds are perhaps the most common basis for sample surveys, there are many other universes which may be studied. People of a certain occupation, for example, may be the subject of study. The Department of Agriculture conducts a number of surveys each year in which it is interested only in farms (38). Often it may define its universe as cotton farmers or soybean farmers, perhaps also restricting the area covered, so that the universe may include only so narrowly defined a population as farmers in the state of Illinois operating farms of more than thirty acres with soybeans as a main crop.

Housewives are a frequently surveyed population, usually in relation to such information as preferences for packaged foods, methods and extent of canning and preserving of foods, incidence and size of home gardens, and the like. Some household information, particularly financial data, cannot be obtained satisfactorily from housewives, and in such cases the universe would be defined as heads of households.

Other occupational classifications have served as the basis for occasional surveys in which the objectives required information from such specialized groups. During World War II, for example, there was a large scale program of survey research in the United States Army, intended to answer a wide variety of practical questions ranging from soldiers' preferences for different articles of clothing to their opinions on the determination of points for their ultimate discharge (42). During the same period there were a number of studies of shipyard workers, steel workers, and war plant workers of various kinds, undertaken to help solve the pressing problems of housing, transportation, absenteeism, and morale which existed in production centers during that period (26).

Many surveys require samples of populations which are distinguished by some common behavior or experience. There have been surveys of veterans, of college graduates, of subscribers to certain magazines, of visitors to state parks, of people who ride trains, and of many other equally specialized kinds of people. Such samples are selected because these people have especial significance in relation to the objectives of the study. For example, a Survey Research Center study designed to assess the factors which influence the purchase of homes drew its sample from a universe of recent home

purchasers people who had only shortly before gone through the experience of selecting a home (47)

Surveys may also base their samples on purely demographic characteristics. A study of Negro reaction to discriminatory practices obviously requires a sample drawn from a universe of Negroes. An analysis of the market for cosmetics is likely to get its information from women rather than men. A survey of the problems attendant on retirement will restrict its sample to people of retirement age. A study of acculturation might require a sampling based on national origin. Or a survey might very well take into account a number of other factors and restrict its universe for example, to white men between the ages of twenty and thirty whose families have been in this country for at least two generations. The determination of the character of the universe depends largely on the objectives of the study.

Sometimes the objectives of a survey call for a universe which cannot be sampled until a prior screening survey has been made. To find a sample of people of recent German descent for example it would probably be necessary to interview a relatively large sample of the general population determining the national origin of each respondent. From this total group those respondents who identified themselves as of recent German origin could be drawn out as a subsample meeting the requirements of the original objectives. This procedure is usually called double sampling and is useful when ever the universe in question has some specific characteristic which is not closely associated with particular localities or is not otherwise identifiable.

In 1948 the Survey Research Center found it desirable to interview a sample of that part of the national population which is relatively well informed regarding international affairs. This was accomplished by drawing from the samples of three previous surveys made by the Center in the area of international affairs those respondents who had shown themselves well informed regarding these issues (44). This sample within a sample was reinterviewed. The application of the technique of double sampling provided a representation of the well informed section of the population.

Survey methods can be applied to the study of a small highly selected universe as well as to broad segments of the population such as those mentioned in the preceding paragraphs. There are many research problems in social science for which only compara

ly few individuals in the total population are relevant. The study of political leadership, for example, would need to concentrate on the infrequent individuals who fulfill the definition of leader. Research into the characteristics of legislative action would have to concern itself primarily with people serving in the role of legislator. An investigation of the behavior of business corporations would require information from those individuals responsible for decisions governing the actions of such corporations. For such studies a cross section survey of the general public would be entirely inappropriate.

A Survey Research Center study of factors affecting industrial mobility provides an example of a specialized survey for which only a relatively few individuals could serve as respondents (25, 46). A major objective of this study was to assess the influences which impel industrial executives to locate their plants at one or another of the various sites available to them. The sample was selected randomly from a list of all the firms located in a certain geographical area (the state of Michigan). In each of the two hundred firms comprising the sample, a major executive was interviewed regarding the advantages and disadvantages of the location of his plant. Decisions regarding relocating plants or selecting sites for expansion are ordinarily made by a few high ranking individuals and they are the only qualified respondents for a study having these objectives.

Although the most impressive uses of the survey technique during the past ten years have been in the sampling of large heterogeneous populations, it seems probable that there will be increasing application of this research method to populations of a more restricted character. As social science develops conceptual schemes based on observations and leading to new observations, it is likely that the survey method (as well as all other available research methods) will be applied more sharply to the study of people of specialized characteristics subject to specifically defined circumstances.

TYPES OF SURVEY DESIGN

Whenever survey data are to be gathered, there must be a decision as to the specific pattern or design which the data collecting will follow. Every scientist attempts to arrange the conditions of his

research so that the data which are forthcoming will bear most effectively on the hypotheses he is attempting to test or the information needs he is trying to fulfill. This is as true of the survey researcher as it is of any other.

The Unweighted Cross Section

The most familiar and the simplest survey design is the single time unweighted cross section. This is the method *par excellence* for the determination of the characteristics of a population at a specific point in time. For example, the systematic selection of every nth card from the register of the undergraduates of ——— College would provide the basis for a description of that body as to age, sex, high school record, college entrance test scores, college grades or any of the other items of information which appear on the sampled cards. Mean scores and distributions could be obtained for each of these characteristics.

The sample data would, of course, also make possible the cross analysis of all of these items so that comparisons of grade averages of students of high and low high school records could be made, correlations between grades and entrance test scores for each college year could be compared, and many other statistical analyses could be carried out.

The principal objective of such correlational analysis is to identify causation through the technique of inference. An example may be found in Cartwright's studies of the relation of personal solicitation of prospective buyers of government bonds during World War II to the actual purchase of these bonds (12). Comparisons of people who had been personally asked to buy bonds with those who had not showed approximately twice as many bond buyers in the first group as in the second. This discrepancy was not reduced when such factors as age, income, education, and place of residence were held constant. The inference is strong that individual purchases of bonds were significantly affected by personal solicitation.

The Weighted Cross Section

A variation of the basic cross section survey design is the weighted cross section. This involves the deliberate oversampling of

some subgroup of the designated universe which has especial importance for the objectives of the survey but is known to be a relatively small fraction of the total population. Thus, in designing a national survey for the study of the uses made of public libraries, the Survey Research Center unbalanced its sample design to bring into the sample a larger number of recent users of libraries than would have been caught in an unweighted cross section (10). This was done by doubling the sampling rate in city blocks having high rentals, on the assumption that people who pay high rents are more likely to visit public libraries. The resulting increase in the number of library users interviewed made it possible to describe the characteristics and habits of this important group with greater confidence than would have been possible otherwise.

Oversampling is especially useful in surveys dealing with the distribution and use of income and savings (43). It is well known that wealth is distributed unequally in this country, some individuals having incomes and assets many times the national average. It is also apparent that in a sample of a few thousand households the influence that a relatively small number of these divergent cases can exercise on the total sample is appreciable. To reduce the sampling error of the data obtained from the top income receivers, and to make possible the analysis of these people as a separate group of the total population, it is customary in such surveys as the Survey of Consumer Finances to increase the total number of high income people by oversampling the areas in which they are most likely to be found. Whenever oversampling of this kind is done, it is necessary, of course, to weight these cases down to their proper contribution to the total sample when the data are analyzed.

Contrasting Samples

It is sometimes more efficient to draw samples from subgroups which contrast in the variable most important to the study than it is to sample the entire universe.

This design is well illustrated in a Survey Research Center study of the influence on public attitudes of the proximity of atomic energy installations (16). The purpose of this study was to find out whether the establishment of atomic energy reactors tended to produce insecurity and apprehension in the surrounding communi-

ties To provide the most effective test of this question, samples were interviewed in several cities situated within twenty five miles of major atomic installations similar samples were interviewed in cities paired with the atomic energy cities in geographical, industrial, racial, and other characteristics but situated at some distance from any major atomic center

A variation of the same design has been used by R. C. Angell in a study of moral integration of cities (1) On the basis of an examination of such indices as homicide rates average contributions to Community Chest, and the like Angell was able to order a list of American cities according to an index of integration He then chose two cities from both the high and low extremes of this scale and interviewed samples from the population of each The purpose of this study was to ascertain the degree to which the factors which had produced the original differentiation of the cities as "well" or poorly integrated were reflected in the way people in these cities evaluated their communities and identified themselves with them

The rationale of the contrasting sample design is that the effects or correlates of a variable thought to be important can be most clearly seen if situations are studied which provide the greatest extremes in the presence of this independent variable Presumably factors which do not vary even under these contrasting conditions are not being influenced by the variable in question As in all such studies in which only the extremes of a distribution are observed, there is danger in assuming that a difference in these extremes reflects a linear relation throughout the total range of the variables considered This, of course, does not necessarily follow

Successive Cross Sections

Studies of change necessarily require measurements at successive points in time In survey research, two types of study utilize the procedure of successive sampling from the same population the before after design and the study of trends

No technique is more common in the total array of research procedures than the before and after measurement of a variable to test the effect of a stimulus, an event, or a change which has been introduced between the first and second measurements In social

science the range of change producing factors which are the subject of study is very broad indeed. Sometimes these factors are manipulated as experimental variables by the researchers, as in the study of the effect of special preschool training on the I Q of young children. Very commonly, however, the social scientist is interested in the effect of events over which he has no control—for example, a declaration of war, an act of Congress, a race riot, the advent of television, or a population movement. Although these events are beyond his control, he can often anticipate their occurrence and arrange his measurements accordingly.

A simple example of the before and after design is found in a Survey Research Center study done in 1946 at the time of the tests of atomic bombings on Bikini atoll (7). The major objective of this study was to measure the effect which this highly publicized event would have on American thinking regarding the atomic bomb, on public anxiety concerning its use, and on popular estimations of its military significance. A nation wide sample of adults was interviewed in June just prior to the Bikini tests. These people were asked a series of questions relating to various aspects of atomic energy and the atomic bomb. In August, after the results of the tests had been widely publicized and discussed, a second sample similar to the first, was interviewed. This second sample was asked the same questions which had been asked in June as well as a number of special questions relating to perception of what the Bikini tests had shown. A comparison of the results of the two surveys showed that, although there was some surprise that the Bikini bombs had not done greater damage, there was no essential change in public perceptions or attitudes regarding the atomic bomb.

A similar design was used by the National Opinion Research Center in a study of the effectiveness of a campaign conducted in Cincinnati intended to educate the population of that city regarding the activities of the United Nations (40). A survey conducted before the campaign established base line scores of public interest in the United Nations, information regarding its activities, and attitudes toward its accomplishments. This study indicated the sections of the population which "showed themselves to be most in need of enlightenment." A comparable sampling six months later, after the campaign, demonstrated the extent to which people had been reached by the campaign and had been influenced by it.

The study of trends differs from the before and after design only in that more than two measurements are taken and the measurements are spaced over a continuing period rather than on either side of a specific event. One of the best known trend studies done by sample surveys is Cantril's study of changes in American attitudes in the period prior to Pearl Harbor toward entry into World War II (11). Through a series of surveys conducted during 1940 and 1941, Cantril was able to trace the gradual rise of public sentiment for aid to England, for resistance to Japanese aggression, and for a declaration of war on Germany.

In a different area of information, the British Survey of Sickness reports each month on prevailing rates of sickness, incapacity and medical consultation. This survey of the British population makes possible the analysis not only of seasonal trends but also of long term changes in the state of public health. In the United States, the Survey of Consumer Finances makes possible the study of annual fluctuations in the economic status of the population.

The same series of surveys may readily serve both to follow trends and to study changes before and after a specific event. This is well illustrated by Logan's analysis of data from the Survey of Sickness in which he compares rates of sickness and medical care during the year prior to the institution of the National Health Service in July 1948 with rates in the succeeding twelve months (33). From this study it was possible to estimate the increase in reported illness (about 5 percent) and medical consultations (about 13 percent) which followed the inauguration of the British system of socialized medicine.

Logan's study also illustrates a more detailed type of analysis which is possible with data from successive surveys. Comparisons through time can be made not only of the total universe represented but also of many subpopulations within the total. Thus Logan was able to show that the increases in illness and medical consultation were greater among women than men, among older than younger people, and among low income rather than high income families.

The Survey of Sickness utilizes a further nicety of design which is especially useful when there is need for a large sample. This is the procedure of overlapping samples. In the Survey of Sickness the

respondent is asked to report on his medical experiences of the previous two months, a sufficiently short length of time to minimize memory error. Each month a new sample of 1000 adults is interviewed. By combining the reports of each two successive months samples for the one month they both report, it is possible to create an effective sample of 8000 reports for each month.

Reinterviews

There are some types of survey objectives which require successive interviews with the same individuals. This may involve only one interview and one reinterview or it may mean a series of interviews extending over months or years.

The reinterview design is used when it is necessary to follow the activities or attitudes of the same individuals through a specified time period. This is illustrated by a study of buying intentions reported by Lansing and Withey in which a national sample of consumers was asked at the beginning of the year whether they planned to buy an automobile or other durable goods during the ensuing twelve months (23, 29). At the end of the year, these same people were interviewed again and were asked whether they had bought a car or other durables during the year past. The objective of this study was to analyze the nature of the planning and decision making preceding a purchase, and this could be done only by following through on the stated plans of specific consumers.

A similar design was used by Campbell and Kahn in their study of the presidential vote in 1948 (8). One of the objectives of this study was to measure the final shifts from stated intentions to vote to the final decision on Election Day. Although the aggregate changes from October to November could have been measured by successive unrelated samples in these two months, individual shifts could not have been analyzed without successive interviews of the same people before and after the election.

In some cases multiple interviews are taken with the same respondents simply because the objectives of the survey are too extensive to be covered in a single interview. In the Indianapolis study of fertility (49), for example, four separate interviews were necessary in each of the sample households to cover all the questions

which had to be asked. In such a situation, there is no advantage in spacing the successive interviews and they will ordinarily be conducted as close together as possible.

Samples which are interviewed repeatedly over an extended period of time are usually referred to as panels (30). The use of the panel design is perhaps best illustrated by the study of the presidential vote conducted in Erie County, Ohio, in 1940 by Lazarsfeld, Berelson, and Gaudet (31). This elaborate project was intended to 'follow the vagaries of the individual voter along the path to his vote and to discover the relative effect of various influential factors upon his final vote. It used a sample of six hundred people who were interviewed seven times in successive months, May through November. By this procedure it was possible to observe individual shifts in inclinations toward one or the other candidate during the campaign and to relate these fluctuations to specific influences or stimuli which led to the change.

Beyond its usefulness for the analysis of factors producing individual change, the panel design has two further virtues. The more obvious of the two is that the same sample interviewed twice is a more sensitive measurement of change than two separate samples of the same universe. This results from the intercorrelation of variables, which is at its maximum in the reinterview design. Moreover, when the same sample is interviewed two or more times, the variations implicit in the conduct of field work tend to be repeated and thus correlated.

The panel design also has the advantage of making possible the description of how the constituency of the various economic and social strata of society changes through time. This type of study would be appropriate, for example, to an analysis of the extent to which the top bracket of income earners in this country is comprised of the same people from year to year. A single survey is sufficient to demonstrate that a certain percentage of the people have incomes over \$10,000 in a specific year, but it would require successive surveys of the same sample to find out what percentage of the people have five year incomes exceeding \$10,000 each year. Some facts can be reliably recalled over a considerable period of time, but for many facts, such as income, the memory error is so great that reporting over more than relatively short periods is quite undependable. This

necessitates repeated additive reports if longitudinal data are to be gathered reliably

An application of this feature of the panel technique is found in the monthly Current Population Survey. This study uses the same respondents for six successive months and can thus estimate not only the total number of unemployed in the nation each month but also the total movement into and out of this group. An increase of 100,000 in the total unemployed group from one month to the next might mean that all the unemployed in the first month remained out of work during the second month and were joined by an additional 100,000 previously employed people. However, it might mean that 300,000 people unemployed in the first month went to work in the second month and their places in the unemployed group were taken by 400,000 people who had been at work during the first month. This total movement of 700,000 people can be followed in the Current Population Survey because changes in the work status of the individual members of the sample are reported month after month.

Two factors act as deterrents to the use of the panel design. The first is the virtually inevitable mortality which occurs in any population sample over even a brief period of time. In a cross section of the national population, a loss of 25 percent or more is to be expected after an interval of one year. A large part of this is due to people's moving from one place to another, some of it to refusals to be interviewed a second time, and the rest to the numerous circumstances which beset the efforts of even the best field organizations. A loss of this proportion does not necessarily result in a biased sample but it creates the possibility of serious bias.

The second serious problem associated with the use of panels is the possibility that the continued interviewing will so sensitize and change the respondents that they are no longer representative of the universe from which they were drawn. In the study of the effect of the Bikini atom bomb tests, for example, it was thought inadvisable to reinterview the before test sample after the test for fear that the first interview would call the attention of these people to the test and stimulate them to follow the news of the event more closely than they otherwise would have. One can easily imagine that respondents who know they are going to be interviewed month after month on

questions of foreign affairs, for example, will consciously or otherwise prepare themselves for the next interview

The effect of these two factors on the data gathered from a panel can be measured by interviewing a control sample, independent of the panel, usually at the end of the panel period. If the data from this fresh sample differ from those of the last panel survey by more than would be expected from errors of sampling, there is reason to believe that reinterviewing has introduced bias

VARIETY OF CONTENT STUDIED

The versatility of the sample survey lies not only in the variety of populations to which it may be applied or in the choice of designs which are available but also in the broad scope of data which may be gathered. Any fact which respondents are able and willing to tell an interviewer may become the subject of a survey study.

Surveys may supply answers to the questions 'how many' and 'how much'. Some such data are available from public records—as, for example, the number of those who voted for a candidate in an election, or the total income of all Americans in a given year. But sample surveys are uniquely qualified to answer such questions as 'How many people believe this country is too aggressive in its foreign policy?' or 'How many people use a public library more than ten times during a year?' or 'How many families own stock in business corporations, and how large is their average holding?'

Such information about the frequency of certain opinions or activities in the total population represents in a certain sense only a preliminary phase of survey research. Surveys are intended primarily to answer the questions of 'who', 'how,' and 'why'. Who are the people who own their homes, or own common stock, or vote Republican—that is, what is the occupational, educational, age etc., distribution of homeowners, stockholders, or Republican voters? How is it that some people have contributed to a Red Cross campaign whereas others have not, have they been solicited, have they known about the Red Cross before, and, if so, what, and what are their attitudes toward the Red Cross? Why do people use a public library—for purposes of additional training, enjoyment, or both? In all these respects, the contribution of surveys is unique and serves

to enrich our store of knowledge—that is, our understanding of what has happened and our ability to predict what will happen

The content of survey questions may be classified in various ways. The following classification divides the total range of questions into four broad areas of content

Personal Data

Surveys often include questions regarding the sex, age, occupation, education, religion, nationality, group membership, and many other personal social characteristics of the respondents. Similarly, they may contain questions about the size of income, assets, debts, and other economic variables. The purpose of these questions is not so much to determine the incidence of these characteristics in the population as it is to provide the basis for the analysis of the relation of sex, occupation, or income to other data obtained in surveys (9, 21, 24)

Environmental Data

In many surveys it is important to know certain facts regarding the circumstances in which the respondents live. These might include data about the character of the local neighborhood, the adequacy of the living quarters, or the proximity of friends or relatives.

A study of library use, for example, would need to determine the relative availability of library services. A study of family relations would probably require information on the location and degree of contact with parents and 'in laws'. A survey of home accidents made by the School of Public Health of the University of Michigan included a detailed investigation of the homes of the respondents—their lighting, floor plan, location of rugs and furniture, condition of stairs, and the like (36). Knowledge of such environmental facts is often needed to explain the behavior which is the survey's principal object of study.

Behavioral Data

Many survey questions deal with the actions or behavior of respondents. In the economic field, for example, spending and saving

(purchases of Government savings bonds, houses, automobiles, television sets, etc.) have been studied in surveys so as to determine the frequency of these activities in a given period and their relation to other activities, and to determine the characteristics of those who have engaged in them. Behavior which is only partly economic is studied in surveys when questions are asked about geographical or occupational movements, vacation trips, or visits to physicians. In the study of political behavior, survey questions about voting, writing to one's Congressman, soliciting party contributions, and the like are relevant.

The analysis of information getting behavior would require questions on newspaper reading, radio listening, television viewing, movie attendance, conversations, and other related activities. The total range of behaviors which might interest a survey planner is obviously very wide.

Level of Information, Opinions, Attitudes, Motives, and Expectations

This broad area of "psychological" data includes many of the most interesting questions available to survey analysis. It is also the area in which there are least likely to be data available from non survey sources.

The determination of level of information is often necessary as background to the study of attitudes or opinions. It is dangerous to assume that issues and events are equally understood by everyone, and it is difficult to assess how people stand unless we know what their understanding of the issues is. A respondent's information level may be measured simply in terms of his awareness or unawareness of an issue or event. For example, does he know of the existence of an organization called UNESCO? Does he know that countries other than the United States participated in the United Nations Korean campaign? Does he know of federal regulations controlling installment credit? Information level may also be measured in terms of the degree of detail the individual possesses. If the respondent knows that there is such an organization as UNESCO, for example, does he know how it is constituted, where it is located, what it does, how it is related to the U N, and so forth?

Questions regarding opinions and attitudes are illustrated by inquiries about what people think are the purposes, achievements, or shortcomings of the United Nations, how they feel about the competence of the federal administration, or whether they prefer a sales tax to an income tax. Attitudes are generalized viewpoints of approval or disapproval. To determine the presence or absence of attitudes and the reasons for holding them—what public issues do what kind of people approve or oppose and why?—is frequently an important survey objective.

Survey analysts are not usually satisfied with obtaining information on people's attitudes regarding specific unrelated public issues, it is often more important to investigate patterns of attitudes and interrelations among different attitudes. In this case we do not ask simply how many and what kind of people approve of the Korean campaign, or of the Atlantic Treaty, or of sending troops to Germany, we want, rather, to determine whether a distinct group of people emerge who approve of all these and other related policy measures in contrast to another group opposed to them. In this way, general attitudes may be discerned and it may be determined which specific attitudes are and which are not influenced by these general attitudes.

The study of motives and expectations represents one of the most challenging areas of survey research. The motive concept stands not only for the stated reasons for behavior—that is the answers to the questions *why* (e.g., *Why did you vote Republican?*)—but *more generally for the forces impelling to action*. Expectations represent the time perspective of a person as it spreads forward into the future—that is his opinions and attitudes about what will happen as well as his intentions and plans.

FORMS OF ANALYSIS

Surveys directed toward a joint study of personal and environmental data, behavior, and attitudes are usually much more productive than those intended to cover only one of these aspects. Among the many possibilities of such integrated inquiries a few important ones will be discussed here.

Comparison of Different Parts of the Sample

Instead of studying the distribution of certain kinds of behavior or of certain attitudes among all people, the sample may be broken into several groups with the purpose of determining the differences in behavior or attitudes among these groups. For example, the analysis of people's perceptions of and actions regarding inflation during 1950-1951 became more meaningful when low-income groups were compared with high-income groups. Similarly, studies have often been made of behavior, opinions, and attitudes of specific educational or age groups. Groups which are compared need not represent demographic classifications but may be real groups to which people belong, such as work groups in factories or offices.

Linking Behavior and Attitudes

In some surveys the most critical analysis requires the comparison of behavioral or attitudinal groups. Attitudes are thus not studied in the abstract but are linked to specific forms of behavior. In a factory, for instance, high-production employees may be separated from low-production employees and their attitudes toward the company or their foreman compared. Or, instead of studying all people's satisfaction or dissatisfaction with public libraries, the survey may contrast the attitudes of those who use the libraries frequently with those who use them rarely or not at all. Contrasting attitudinal groups have been established by asking people whether they feel that they are financially better or worse off than a year ago. Such a division of American families has made it possible to study whether purchasing of automobiles is or is not associated with the feeling of being better off financially (22, 23, 28).

The Study of Motivations

In most surveys the direct question of why a certain action was taken represents one, but not the only, approach to the analysis of motivational forces. Since many aspects of the prevailing motivational forces may not be salient in people's minds or may not be recognized by them as having contributed to a decision, it is necessary to use a correlational approach. One example of

unearthing motivational forces though correlation studies has been mentioned earlier. When people were asked during World War II why they had bought war bonds, they gave many reasons (usually patriotic or investment reasons) but rarely mentioned that they had purchased bonds because they had been solicited to do so. A comparison of people with similar incomes and occupations who had and who had not been solicited revealed, however, that many more of the former than of the latter had purchased bonds.

Similarly, inquiries about reasons for purchasing automobiles may shed light on many important factors which have at one time or the other contributed to the decision to buy a new car. When asked why they bought a car during the preceding year, very few people mention that their income had increased. However, comparison of car buying of those who had an income increase with those who did not have such an increase reveals the existence of such a relationship (23).

Another example is found in the analysis of factors contributing to investment decisions (building of new plants, purchasing new equipment) on the part of manufacturers. In a survey of top executives or owners of manufacturing plants, the question of why expansion was undertaken or expansion plans were entertained yielded a great variety of reasons. Further relevant factors were discerned, however, by asking the manufacturers questions about their profit expectations and by studying the relation of their answers to the presence or absence of expansion plans. The joint use of correlation analysis and of the direct question of "Why?" proved fruitful in helping to explain why some manufacturers entertained expansion plans though they expected stable or decreasing profits (25).

Making Predictions

Although associations or correlations between different variables do not necessarily show which is the cause and which the effect, such studies may provide information useful for the study of causation. If relationships are established, for example, between past income increases and purchases of durable goods, a prediction about future behavior in case of widespread and substantial income increases may be made even if the question about causation has

Comparison of Different Parts of the Sample

Instead of studying the distribution of certain kinds of behavior or of certain attitudes among all people, the sample may be broken into several groups with the purpose of determining the differences in behavior or attitudes among these groups. For example, the analysis of people's perceptions of and actions regarding inflation during 1950-1951 became more meaningful when low income groups were compared with high income groups. Similarly, studies have often been made of behavior, opinions, and attitudes of specific educational or age groups. Groups which are compared need not represent demographic classifications but may be real groups to which people belong, such as work groups in factories or offices.

Linking Behavior and Attitudes

In some surveys the most critical analysis requires the comparison of behavioral or attitudinal groups. Attitudes are thus not studied in the abstract but are linked to specific forms of behavior. In a factory, for instance, high production employees may be separated from low production employees and their attitudes toward the company or their foreman compared. Or, instead of studying all people's satisfaction or dissatisfaction with public libraries, the survey may contrast the attitudes of those who use the libraries frequently with those who use them rarely or not at all. Contrasting attitudinal groups have been established by asking people whether they feel that they are financially better or worse off than a year ago. Such a division of American families has made it possible to study whether purchasing of automobiles is or is not associated with the feeling of being better off financially (22, 23, 28).

The Study of Motivations

In most surveys the direct question of why a certain action was taken represents one but not the only approach to the analysis of motivational forces. Since many aspects of the prevailing motivational forces may not be salient in people's minds or may not be recognized by them as having contributed to a decision, it is necessary to use a correlational approach. One example of

unearthing motivational forces though correlation studies has been mentioned earlier. When people were asked during World War II why they had bought war bonds, they gave many reasons (usually patriotic or investment reasons) but rarely mentioned that they had purchased bonds because they had been solicited to do so. A comparison of people with similar incomes and occupations who had and who had not been solicited revealed, however, that many more of the former than of the latter had purchased bonds.

Similarly, inquiries about reasons for purchasing automobiles may shed light on many important factors which have at one time or the other contributed to the decision to buy a new car. When asked why they bought a car during the preceding year, very few people mention that their income had increased. However, comparison of car buying of those who had an income increase with those who did not have such an increase reveals the existence of such a relationship (23).

Another example is found in the analysis of factors contributing to investment decisions (building of new plants, purchasing new equipment) on the part of manufacturers. In a survey of top executives or owners of manufacturing plants, the question of why expansion was undertaken or expansion plans were entertained yielded a great variety of reasons. Further relevant factors were discerned, however, by asking the manufacturers questions about their profit expectations and by studying the relation of their answers to the presence or absence of expansion plans. The joint use of correlation analysis and of the direct question of "Why?" proved fruitful in helping to explain why some manufacturers entertained expansion plans though they expected stable or decreasing profits (25).

Making Predictions

Although associations or correlations between different variables do not necessarily show which is the cause and which the effect, such studies may provide information useful for the study of causation. If relationships are established, for example, between past income increases and purchases of durable goods, a prediction about future behavior in case of widespread and substantial income increases may be made even if the question about causation has

not been entirely clarified. Detailed inquiries about the circumstances in which behavior has taken place may result in statements which imply information on causation as: People who expect substantial increases in income will save relatively little.

Asking people about their plans and intentions provides another method of deriving predictions of things to come. It must be emphasized, however, that the translation of expressed plans into predictions is by no means a simple process. If a survey determines for example that a proportion of all American families representing three million people plans to buy a new car in a given year, the prediction that three million new cars will be bought in that year is not justified. Plans are subject to change with circumstances and purchases may be made by people who at an earlier time did not plan to make them. If, however, two consecutive surveys find that the number of prospective automobile buyers has increased from three to four million from one year to the next, the statement that the automobile market is firmer in the second year than in the first may be justified. Expressed plans represent attitudes prevailing at a given time and information about them increases our knowledge about the situation at that time. The greater our knowledge, the better is our ability to predict (23-28-29).

In some cases survey data can be used very effectively to predict public reaction to events which are known to be forthcoming. A survey among industrial workers at the end of 1942, for example, showed that many of these people had no cash or other assets except government bonds (48). Since the survey also showed that many of these workers were making incomes on which they would have to pay income tax, it was easy to foresee that in March 1943 they would have to cash some of their government bonds to meet their income tax obligations. A series of predictions of this kind regarding public purchase and redemption of war bonds was made during World War II by a program of survey research sponsored by the U. S. Treasury Department.

VARIETY OF FIELDS OF APPLICATION

From the diversity of the examples of survey content presented in the foregoing pages, it is apparent that the survey method is

applicable to various fields and scientific disciplines. It may be argued that certain surveys belong in the realm of psychology, others in sociology, still others in economics, or political science, or public health. Although such classifications may be justified from a certain point of view, it must be emphasized that the survey method is essentially interdisciplinary or, to put it more accurately, that it contributes to the integration of several traditionally separate disciplines.

Among surveys which are primarily in the domain of social psychology or sociology are investigations of group belonging, of leader-follower relations, of family life, of occupational choice. Studies of income, expenditures, and savings may be classified as economic surveys. Studies of voting, of participation in political movements, and of the distribution of political attitudes may be thought to belong to political science. Surveys of the incidence of illness, uses of medical services, or the nature of public beliefs regarding health concern the field of public health.

Such classifications, however, are somewhat arbitrary. All surveys, even those in economics or public health, have something to do with people's behavior. If we ask about level of information, opinions, or attitudes, we are primarily interested in finding out how and why people behave as they do. If we omit enumerative surveys of the census type (intended, for example, to determine the number of employed workers or the number of farmers raising wheat in the United States), we may conclude that all surveys have some relation to psychology. They are concerned fundamentally with people's behavior—their social behavior, their economic behavior, their political behavior, their health behavior. Although statistical reliability requires the grouping of individuals, survey data always derive from individual reports. Finally, psychology enters into the picture because of the method of surveys: the contact between respondent and interviewer represents an interpersonal relationship of the kind that has historically interested professional psychologists.

It is not permissible, however, to classify survey research as a part of psychology. Survey research has no specific disciplinary anchor point. It is being used by specialists in all fields of behavioral science, being adapted in each case to the requirements of that field. Survey data are broadening the empirical base of a variety of

fields. They are also providing the raw materials for an increasing volume of cross disciplinary analysis, which, it may be hoped, will serve in time to help bring about a closer integration of the presently separate behavioral sciences.

Depending on the intent of the survey planner, surveys can be a tool of *applied* research or can have a function in what is considered *basic* research. When data are collected which policy makers in business or government need for their immediate practical purposes, it is customary to speak of applied research. The survey technique has been widely used by businessmen in their research on markets, consumer preferences, buying habits, and the like. Numerous large commercial research agencies are continually engaged in surveys of this kind for American business. Various branches of the federal government have also found reason to conduct applied research of this kind, usually to assess public response to their programs and to find ways in which the programs can be improved (4).

The use of the survey method as a basic tool of the behavioral sciences has been discussed in the preceding section. A great variety of hypothesis testing has been done through surveys, for example, in studies of the relation of economic experiences and expectations to spending and saving behavior (22, 23), the relation of personal frustration to aggression against minority groups (5), the relation of identification with political parties to attitudes regarding political issues (2) and the relation of distance to the transmission of rumor (14).

The distinction between basic and applied research, though clear-cut in some instances, often proves to be superficial. What is called basic research may have more significant practical uses than a great deal of applied research. Suppose the survey technique is used to study the dynamics of inflation—as it is, in fact, being used. If reliable information were available about the factors which induce businessmen and consumers to stock up or to buy in advance and in excess of their needs, our basic knowledge of economic behavior would be greatly enhanced. At the same time, practical applications regarding anti inflationary policies would emerge from such findings. The discussion of the probable effects of new policy measures could eventually be removed from the sphere of hunches and guesses and could be based on scientific evidence.

Similarly, the extensive surveys conducted among members of

the armed forces during World War II may be regarded, from one point of view, as applied research. Some of these surveys—for example, those concerned with the attitudes of and toward Negro soldiers—were intended to assist policy decisions regarding the integration of Negro and white troops (42). At the same time, these studies can correctly be regarded as basic research, since they serve to clarify the possibilities of changing attitudes through personal experience and manipulation of the environment.

FLOW CHART OF A SURVEY

The sequence of the tasks involved in carrying out a survey, from the first stages of planning to the preparation of the final report, is presented in this section.

1. *General Objectives*

The problems which make a survey necessary and the general objectives of the survey are stated. This statement is usually expressed in broad terms and defines only the general area and scope of the project.

2. *Specific Objectives*

Although the general objectives, usually few in number, are formulated without regard to the requirements of the survey technique, these are considered when the general objectives are broken down into the usually numerous specific objectives. The specification of all data to be gathered and of the hypotheses to be tested by the survey is accomplished at this stage.

3. *Sample*

Two decisions must be made regarding the survey sample: (1) what the universe of the survey is to be (will it be all American families, or all employees of a factory, or all the physicians living in a region of the country?) and (2) the size and design of the sample which is to be drawn. After these decisions are made, the

actual drawing of the sample units and the preparation of delineated maps, block lists, and the like may proceed

4 Questionnaire

The method by which the sample is to be contacted (by personal interview, telephone, or mail) is determined, and a questionnaire is prepared. The questionnaire is not simply a translation of the specific objectives into language understandable to the respondents, it is built carefully, with regard to the type of questions to be asked, the degree of probing, the sequence of the questions, and the establishment of rapport. The draft of the questionnaire is pretested in the field before its actual use.

5 Field Work

When personal interviews are to be conducted, interviewers must be trained both in general interviewing procedures and in questions specific to a given survey. Interviewers are supplied with an instruction manual which explains the objectives of the study and the meaning of each question. Provision is made for careful supervision of the interviewing.

6 Content Analysis

The data obtained in a survey may be so simple that the interviews received may be easily and directly transcribed into tabulations (or into punched cards through which tabulations are made). But even surveys of the census type require careful editing, and attitude and opinion surveys require content analysis. This is done by preparing a code, a numbered list of major items subsuming all the responses received to each question. Coders must be trained, and coding must be supervised and its reliability established.

7 Analysis Plan

The questionnaire of a large scale survey may contain 50-100, or more questions. It would be very inefficient to tabulate the relationship between the responses received to each question. The analy-

sis plan, which, in case of the use of machine tabulating equipment, results in writing out "machine requests," contains the machine runs which are needed to test the hypotheses enumerated in Step 2. This plan has been implicit in the surveyor's thinking from the very beginning of the study. His anticipation of the tabular material necessary to answer the objectives of the survey underlay the preparation of the questionnaire and the determination of the content analysis.

8. Machine Tabulations

The results of the coding process are used to prepare punched cards, and the tabulations foreseen in Step 7 are carried out.

9. Analysis and Reporting

The data are analyzed, their reliability is determined, and a report is written embodying the survey findings. Sometimes, in the case of administrative or applied studies, survey findings are used as the basis of conferences with policy-makers for the interpretation of the implications of the research data for action decisions. This type of reporting is sometimes called "feedback."

This scheme of the sequence of survey work should not imply that the nine steps are independent of one another. Some of the steps listed in succession are usually carried out simultaneously—for example, the code may be prepared before the field work, and content analysis may be carried out at virtually the same time as interviewing. The survey process is a highly interconnected chain of events. Decisions at each step must be congruent with what has gone before and must anticipate what will follow (32, 34).

RELIABILITY AND VALIDITY

The reliability of survey data can be measured in the same way as the reliability of any other kind of research data, by retest. Both individual and aggregate scores can be analyzed in this way.

Unreliability in survey data results from a combination of dif-

actual drawing of the sample units and the preparation of delineated maps, block lists, and the like may proceed

4 Questionnaire

The method by which the sample is to be contacted (by personal interview, telephone or mail) is determined, and a questionnaire is prepared. The questionnaire is not simply a translation of the specific objectives into language understandable to the respondents, it is built carefully, with regard to the type of questions to be asked, the degree of probing, the sequence of the questions, and the establishment of rapport. The draft of the questionnaire is pretested in the field before its actual use.

5 Field Work

When personal interviews are to be conducted, interviewers must be trained both in general interviewing procedures and in questions specific to a given survey. Interviewers are supplied with an instruction manual which explains the objectives of the study and the meaning of each question. Provision is made for careful supervision of the interviewing.

6 Content Analysis

The data obtained in a survey may be so simple that the interviews received may be easily and directly transcribed into tabulations (or into punched cards through which tabulations are made). But even surveys of the census type require careful editing, and attitude and opinion surveys require content analysis. This is done by preparing a code—a numbered list of major items subsuming all the responses received to each question. Coders must be trained, and coding must be supervised and its reliability established.

7 Analysis Plan

The questionnaire of a large scale survey may contain 50–100 or more questions. It would be very inefficient to tabulate the relationship between the responses received to each question. The analy

sis plan, which, in case of the use of machine tabulating equipment, results in writing out "machine requests," contains the machine runs which are needed to test the hypotheses enumerated in Step 2. This plan has been implicit in the surveyor's thinking from the very beginning of the study. His anticipation of the tabular material necessary to answer the objectives of the survey underlay the preparation of the questionnaire and the determination of the content analysis.

8 *Machine Tabulations*

The results of the coding process are used to prepare punched cards, and the tabulations foreseen in Step 7 are carried out.

9 *Analysis and Reporting*

The data are analyzed, their reliability is determined, and a report is written embodying the survey findings. Sometimes, in the case of administrative or applied studies, survey findings are used as the basis of conferences with policy makers for the interpretation of the implications of the research data for action decisions. This type of reporting is sometimes called 'feedback'.

This scheme of the sequence of survey work should not imply that the nine steps are independent of one another. Some of the steps listed in succession are usually carried out simultaneously—for example, the code may be prepared before the field work and content analysis may be carried out at virtually the same time as interviewing. The survey process is a highly interconnected chain of events. Decisions at each step must be congruent with what has gone before and must anticipate what will follow (32, 34).

RELIABILITY AND VALIDITY

The reliability of survey data can be measured in the same way as the reliability of any other kind of research data, by retest. Both individual and aggregate scores can be analyzed in this way.

Unreliability in survey data results from a combination of dif

ferent types of error. Interviewing error arises from inconsistencies in the way in which the interview is conducted. Reporting error results from vagaries of mood or attitude on the part of the respondent. Sampling error is implicit whenever a sample is taken as representative of a universe. Errors in coding, tabulation, and analysis make their inevitable contribution to the total. Anything which tends to create different results under theoretically identical conditions may be said to be contributing to unreliability.

The reliability (or consistency) of the information given by an individual respondent can be assessed either through related questions in the same interview or through the same question in successive interviews. In the former case, for example, the reliability of the reporting of personal financial data can be estimated by balancing the income and expenditure data which are given in individual interviews. In the same ways respondents' reports of their age can be checked against their report of their educational and employment history. Questions intended to measure degree of information can be ordered in a scale of difficulty and the extent of inconsistency of response within individual interviews noted. The reliability of attitudinal responses can be assessed in a similar way, using scaling procedures of the type developed by Guttman (41). Such measures involve some ambiguity, since they reflect not only individual unreliability but also failure to achieve unidimensional scales.

The measurement of the reliability of report by the comparison of individual responses in successive interviews involves special problems. If the two interviews follow each other within a very short time interval, it is easily possible that the respondent will remember his earlier answers and simply repeat them verbatim in order to appear consistent. On the other hand, if a long time interval elapses, error may be introduced by the respondent's inability to remember the data he is asked for. This latter problem becomes aggravated if the datum which the respondent is asked to report is, like income for a specific year, a point in a changing series.

It has been shown that some items—such as age, religion, and country of origin—are reported with a very high degree of consistency (over 90 percent identity) over periods of a year or more (6). It has also been shown, however, that reports of annual income made a year after the end of the reported year often depart sub-

stantially from the report given immediately following the reported year and that these deviations tend to be biased in the direction of the individual's income change during the year following the original report (50)

Memory errors of this kind are so pervasive that there is a strong tendency in survey studies to reduce to a minimum the period the respondent is asked to recall. Income data are usually requested on a yearly basis, and preferably shortly after the end of the calendar year, when income tax obligations compel most people to total up their year's earnings. The Survey of Sickness, as we have seen, uses a reporting period of two months. Surveys of radio listening customarily ask for a report of only the previous day. Generally speaking, the more ephemeral and less eventful the experience to be reported, the shorter the reporting period.

The reporting error of attitudinal responses is difficult to assess in successive surveys because of the problem of differentiating true change which may have taken place between surveys from simple inconsistency of report. For example, the same question (Considering the country as a whole, do you think we will have good times or bad times or what during the next 12 months?) was asked twice, at the beginning of 1948 and at the beginning of 1949. From a sample of 655 identical respondents, it was found that 41 percent gave the same answers and 18 percent gave radically different answers (22). The distribution of the aggregate scores was quite similar at the two dates. Nevertheless, some of the consistency (41 percent) may have been due to chance and some of the change (18 percent) may not have been true change. There has not been sufficient research reported to justify any generalizations as to whether certain types of attitudinal responses are more or less persistent than others or are reported with greater or less reliability by individual respondents.

It is customarily found that the consistency of averages or frequency distributions is greater than that of individual scores. In Withey's study, for example, the distributions of 1947 urban incomes were found to be quite similar in surveys conducted early in 1948 and early in 1949 (the same respondents were interviewed in both surveys), although the individual scores which make up the distributions were far from perfectly correlated as Table I indicates (50).

TABLE I

Comparison of Income Data Obtained in Two Successive Surveys

1947 Money Income Before Taxes	From Survey Conducted in		Proportion in Same Bracket in Both Surveys	Proportion in Indicated Bracket in First Survey	
	Early 1948	Early 1949		Adjacent Bracket in Second Survey	Nonadjacent Bracket in Second Survey
Under \$1 000	8%	7%	6%	1%	0%
\$1 000-\$1 999	14	14	10	4	•
2 000- 2 999	23	28	17	5	1
3 000- 3 999	22	18	12	9	2
4 000- 4 999	14	13	6	6	2
5 000- 7 499	12	13	7	4	1
7 500 and over	7	7	6	1	•
	<u>100%</u>	<u>100%</u>	<u>64%</u>	<u>30%</u>	<u>6%</u>

* Less than one half of 1 percent

SAMPLE 415 identical urban spending units who were interviewed once in early 1948 and once in early 1949 and who both times gave information about their 1947 income

SOURCE Surveys of Consumer Finances conducted by the Survey Research Center for the Federal Reserve Board

The reliability of frequency distributions from comparable but independent samples can be readily demonstrated by splitting the total sample of a survey into randomly selected subsamples. In Table II the reported income of some 1200 respondents interviewed in a national survey of attitudes toward big business is compared to the distributions obtained from combining the reports of every fourth respondent (17). The consistency which appears is not uncharacteristic of survey data based on careful sampling and interviewing methods.

The high degree of comparability of distributions from successive surveys can be demonstrated not only for demographic data such as income or education but also for psychological data. This is well illustrated by Cartwright's data on bond buying (13). The reasons people gave for the Government's interest in selling bonds during World War II occurred in very similar proportions in one survey after another (Table III).

TABLE II
Income Distributions of Systematically Selected
Subsamples of a National Sample

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>Total</i>
\$ 0-\$ 999	11%	8%	10%	9%	9%
1,000- 1,999	11	13	12	9	11
2,000- 2,999	23	18	19	21	20
3 000- 3,999	19	19	20	20	20
4,000- 4,999	11	14	11	13	12
5,000- 7,499	12	15	16	14	14
7,500- 9,999	5	4	4	3	4
10 000 and over	4	3	4	4	4
Don't know	1	2	1	2	2
Not ascertained	3	4	3	5	4
	<hr/> 100%	<hr/> 100%	<hr/> 100%	<hr/> 100%	<hr/> 100%
Number of interviews	317	308	299	303	1,227

TABLE III
Reasons Attributed to Government for Wanting to Sell Bonds

<i>Reasons</i>	<i>Jan</i> <i>1944</i>	<i>June</i> <i>1944</i>	<i>Nov</i> <i>1944</i>	<i>June</i> <i>1945</i>
To finance the war, to win the war, to help soldiers	65%	65%	67%	68%
To prevent inflation	14	15	15	14
To get people to save	7	8	7	10
To provide postwar security	2	3	2	3
Other reasons	12	9	9	5
	<hr/> 100%	<hr/> 100%	<hr/> 100%	<hr/> 100%
Number of interviews	1 441	1 925	2 148	2 263

Despite the many opportunities for errors to enter the survey process, there is no doubt that when surveys are conducted with proper observation of the basic tenets of research, many types of data can be collected with not only tolerable but reassuring reliability

The validation of survey data often presents serious problems. The customary procedure for establishing the validity of measurements made in social research is through a comparison with an outside criterion. Unfortunately, there is not always an acceptable criterion available when survey data are gathered. The survey is likely to have been done precisely because there were no relevant data at hand.

Validation on an individual basis can be achieved by direct comparison of the information given by individual respondents with records available from other sources. In a study reported by Cahalan, examination of courthouse records, automobile registrations, and other official documents showed a high correspondence with the report of individual respondents, although some items showed greater discrepancy than others (3). Not all differences in such a comparison can be regarded as survey inaccuracies, because official records are often not entirely up to date or complete and may not yet show changes which are reported in the survey.

A similar kind of validation can be made when respondents of known characteristics are selected as the sample and they are subsequently interviewed regarding the characteristics in question. Hyman has reported such a study in which a sample of people who had cashed government bonds during World War II were interviewed to determine the uses to which this money was put (20).

Seventeen percent of these people denied that they had cashed any bonds. This relatively high discrepancy reflects the problem encountered when people are asked to reveal information which is not entirely to their credit.

A common method of demonstrating survey validity consists of comparing survey distributions with comparable distributions from the preceding Decennial Census. Virtually every survey reports distributions of the demographic characteristics of its sample—age, sex, race, education and occupation—and comparisons with Census data can, at the least, illustrate the absence of gross errors.

A more rigorous test of survey validity consists in comparing aggregate data derived from sample surveys with independent outside estimates of the same magnitudes. Unfortunately, however, relatively few aggregate data derived from surveys are directly comparable with outside estimates and, most commonly, the comparison requires complex adjustments for differences in concepts or coverage. How

ever, it is possible, for example, to check the validity of the number of births reported by a sample of families for the preceding year by comparing this figure with the national birth rate as compiled from official records of the Office of Vital Statistics. Or the number of automobiles reported owned by a sample can be expanded to a national total and checked against automobile registration figures.

The most detailed comparisons of this kind are those of income data. Selma Goldsmith has reported comparisons of aggregate annual incomes, derived from several sample surveys conducted by different organizations, with the national income as estimated by the U. S. Department of Commerce primarily from available records such as payroll statistics (18). In these comparisons, which required very complex adjustment procedures, the discrepancies from criterion figures were found to be relatively small, mostly within the range of the sampling error.

Although many examples could be given of reassuring comparisons of survey data with outside figures, it must be recognized that large discrepancies are also occasionally found, despite observance of the most rigorous standards of survey methodology. In some cases it seems virtually impossible to overcome reporting error, which may result in very serious bias. This is relatively uncommon, however, and is usually associated with deliberate falsification (as in the report of currency on hand) or with memory errors (as in the report of amount of bank deposits held year before last).

The validation of data on attitudes, expectations, intentions, and the like presents problems of a different kind. In this case there is no outside "true score" which can be taken as a criterion. The statement which the respondent makes at the time of the interview is in itself a datum for analysis. It may or may not correlate with other nonverbal behavior. Its usefulness for predictive purposes is greater if it does, but its intrinsic validity is not dependent on this. Invalidity is present of course if the respondent consciously or otherwise misrepresents his attitudes or intentions because of unwillingness to cooperate with the interviewer. Such misrepresentation is often very difficult to detect.

Despite the absence of outside criteria, there are many internal tests which can be applied to attitudinal data to demonstrate that they are more than expressions of idle caprice on the part of the respondents. One finds, for example, that approval of proposals for

the government to provide jobs for everyone is much higher among low income people than among high, that intentions to buy durable goods are higher among people who expect increases in income than it is among those who expect decreases, that expressions of hostility toward Negroes are more common among people in Southern cities than among people in Northern cities, that people who disapproved of the Taft Hartley Act in 1948 were much more likely to vote Democratic than those who approved it. These findings are consistent with our general expectations based on other information about our society and its functioning. This is by no means as convincing a validation of survey data as an established outside criterion would provide. In the absence of such criteria, however, analysis of the internal logic of survey data can often present an impressively consistent picture.

When survey data are used for purposes of prediction, additional questions of validity are encountered. A respondent's statement in October of his intention to vote may be a valid expression of his inclination at that time but an invalid indication of his actual vote in November. A consumer's expression in January of his intention to buy a car during the ensuing twelve months may be a true representation of his intent, but there are many unforeseen contingencies which may prevent him from carrying it out. The fact that there is not exact correspondence between intentions expressed by individual respondents and their subsequent actions does not mean, of course, that trend data derived from repeated surveys of intentions cannot be of value. For example, the trend of intentions to buy durable goods (expressed as proportions of the total sample) determined in the annual Survey of Consumer Finances has proved to be indicative of the trend of subsequent purchases. Schweiger has shown, also, that this correspondence of trends holds good not only for the population at large but also for various subgroups of the population, for high income people or skilled workers, for example (37).

LIMITATIONS

The foregoing pages have illustrated the scope of the survey technique. It is clearly a research instrument of great versatility,

applicable to a wide range of problems in the general area of social research. It has evident limitations, however, both in precision and adaptability.

It is apparent that any data gathering project based on a sample is subject to sampling error. This means that all findings coming from such a study must be interpreted in the light of this error. This limitation becomes particularly important when a total sample is divided into parts for purposes of analysis. Every researcher who has dealt with data from sample surveys has experienced the frustration of running out of cases as he pursued his analysis into smaller and smaller cells. This is especially true when the analysis calls for division of the sample into geographic regions. Few samples are large enough to permit regional analysis on any but the broadest basis.

Survey research is also subject to all of the errors of measurement implicit in any attempt to derive a score representing a person's attitudes, abilities, behaviors, or other traits. Since surveys usually depend on the voluntary cooperation of respondents, it is often not feasible to use the tedious psychophysical methods which might seem desirable for the greatest reduction of measurement error. Compromises of this kind may not be necessary when the subjects of study can be coerced (college students or military personnel, for example), but they may well be when the respondents are randomly chosen adults in their own homes.

Despite the large number of questions that can be asked in a single survey, there are limits to the number of topics that can be covered. Every data gathering instrument has an optimal length for the population to which it will be submitted. Beyond that point, interest begins to lapse and cooperation to diminish. The survey researcher must take care that he does not overestimate the tolerance of his respondents, and he commonly finds that this involves sacrificing questions to which he would like very much to have answers.

The limitation on length may be especially important in areas which are thought to require a long period of questioning before rapport can be established and resistance overcome. This might apply, for example, in the study of personal maladjustment or family discord, where clinical experience might lead one to expect that a first interview of an hour's length would not be sufficient to uncover the underlying problem. This seems a reasonable assumption.

the government to provide jobs for everyone is much higher among low income people than among high, that intentions to buy durable goods are higher among people who expect increases in income than it is among those who expect decreases, that expressions of hostility toward Negroes are more common among people in Southern cities than among people in Northern cities, that people who disapproved of the Taft Hartley Act in 1948 were much more likely to vote Democratic than those who approved it. These findings are consistent with our general expectations based on other information about our society and its functioning. This is by no means as convincing a validation of survey data as an established outside criterion would provide. In the absence of such criteria, however, analysis of the internal logic of survey data can often present an impressively consistent picture.

When survey data are used for purposes of prediction, additional questions of validity are encountered. A respondent's statement in October of his intention to vote may be a valid expression of his inclination at that time but an invalid indication of his actual vote in November. A consumer's expression in January of his intention to buy a car during the ensuing twelve months may be a true representation of his intent but there are many unforeseen contingencies which may prevent him from carrying it out. The fact that there is not exact correspondence between intentions expressed by individual respondents and their subsequent actions does not mean of course that trend data derived from repeated surveys of intentions cannot be of value. For example the trend of intentions to buy durable goods (expressed as proportions of the total sample) determined in the annual Survey of Consumer Finances has proved to be indicative of the trend of subsequent purchases. Schweiger has shown also that this correspondence of trends holds good not only for the population at large but also for various subgroups of the population for high income people or skilled workers for example (37).

LIMITATIONS

The foregoing pages have illustrated the scope of the survey technique. It is clearly a research instrument of great versatility

applicable to a wide range of problems in the general area of social research. It has evident limitations, however, both in precision and adaptability.

It is apparent that any data gathering project based on a sample is subject to sampling error. This means that all findings coming from such a study must be interpreted in the light of this error. This limitation becomes particularly important when a total sample is divided into parts for purposes of analysis. Every researcher who has dealt with data from sample surveys has experienced the frustration of running out of cases as he pursued his analysis into smaller and smaller cells. This is especially true when the analysis calls for division of the sample into geographic regions. Few samples are large enough to permit regional analysis on any but the broadest basis.

Survey research is also subject to all of the errors of measurement implicit in any attempt to derive a score representing a person's attitudes, abilities, behaviors, or other traits. Since surveys usually depend on the voluntary cooperation of respondents, it is often not feasible to use the tedious psychophysical methods which might seem desirable for the greatest reduction of measurement error. Compromises of this kind may not be necessary when the subjects of study can be coerced (college students or military personnel, for example), but they may well be when the respondents are randomly chosen adults in their own homes.

Despite the large number of questions that can be asked in a single survey, there are limits to the number of topics that can be covered. Every data gathering instrument has an optimal length for the population to which it will be submitted. Beyond that point, interest begins to lapse and cooperation to diminish. The survey researcher must take care that he does not overestimate the tolerance of his respondents, and he commonly finds that this involves sacrificing questions to which he would like very much to have answers.

The limitation on length may be especially important in areas which are thought to require a long period of questioning before rapport can be established and resistance overcome. This might apply, for example, in the study of personal maladjustment or family discord, where clinical experience might lead one to expect that a first interview of an hour's length would not be sufficient to uncover the underlying problem. This seems a reasonable assumption.

tion, although the Kinsey studies of sexual behavior indicate that under some circumstances a short interview can bring out personal information which is ordinarily carefully concealed (27)

A sample survey designed to represent a population dispersed over a wide geographical area is likely not to give an adequate representation to any population characteristic which is highly localized. This means that the influence of specific local social groups, for example, cannot be assessed through the usual national survey, since it is unlikely that more than a very few members of any such group would be caught in a cross section sample. The study of local community factors requires a concentration of effort on the specific community rather than the broad dispersion which is desirable when a widely scattered population is to be represented.

It is impossible to analyze adequately the complex fabric of social organization through the survey method alone, because the process of sampling tends to lift the individual respondent out of his social context. Other methods are better adapted to the study of the countless interconnections which give society its integration. It is apparent that the survey method is not well suited to studies of historical development. Ordinarily, survey reporting refers to a specific point in time or to a relatively short time period. Studies of origins and long term developments require research methods of a more longitudinal character.

The most obvious limitations of the survey procedure arise from the fact that it almost inevitably requires a considerable investment of manpower and time. Small scale surveys of highly localized and accessible populations can, of course, be successfully carried out by a single individual, assuming he has the requisite skills and diligence. If time is not a pressing consideration, a small group of researchers can carry through projects of considerable proportions, as evidenced by the Kinsey survey. More commonly, however, surveys are conducted by groups of social science technicians, sometimes several hundred on a single project. This may include specialists in study design, sampling questionnaire construction, interviewing, coding, machine tabulation and statistical analysis. The technology of surveying has become so complicated that professional training in this field (or at least advice) is virtually mandatory if the many pitfalls which await the untrained are to be avoided.

The ponderous character of many surveys imposes a further restraint on the researcher which he may find irksome. Once the design of a survey is set the survey must be carried through according to those specifications. This means that it may take months before a specific hypothesis can be tested and each new variation of the study design intended to carry the theoretical development further will require additional months. To the laboratory researcher accustomed to varying his experimental design every month or so this slow pace may seem an intolerable frustration.

The full contribution of survey research to the development of the behavioral sciences however can be achieved only through continuing programs of research extending over a period of years. The value of such programs is not primarily in their ability to repeat the same observations at different points in time. Much more important is the opportunity they provide for the application of an integrated framework of theory to diverse aspects of behavior and for progressive revision and improvement of the hypotheses tested.

CONCLUSION

Although the origins of the sample survey go back to the nineteenth century, the advanced methodologies now available to survey researchers have been developed only within the last few years. None of the specific studies referred to in this chapter was done before 1940, few of them could have been done before that time.

The experience of the past ten to fifteen years has demonstrated to social scientists generally something of the potentialities and limitations of the sample survey. There can be little argument that this development has given an important stimulus to the quantitative study of social phenomena. There is reason to believe that the possibilities of the survey technique are still not fully realized either in scope of applications or in precision of methods.

BIBLIOGRAPHY

-
- 1 Angell R C The moral integration of American cities *Amer J Sociol*, 1951 47, No 1 Part 2, 1140
 - 2 Belknap G and Campbell, A Political party identification and attitudes toward foreign policy *Publ Opin Quart*, 1951 52, 15 No 4, 601 623
 - 3 Cahalan D *Validity of behavior reports in opinion surveys* A paper presented at Amer Stat Assoc Meetings December 1949
 - 4 Campbell A The uses of interview surveys in federal administration *J Soc Issues*, 1946 2, No 2 14 22
 - 5 ——— Factors associated with attitudes toward Jews In Newcomb T M and Hartley E L (ed) *Readings in social psychology* New York Holt 1947 pp 518 527
 - 6 ——— *Attitude stability and change a re-interview study of the national population* A paper presented at Amer Psychol Assoc Meetings September 1948
 - 7 ——— Eberhart S and Woodward P *Public reactions to the atomic bomb and world affairs Part II Findings of the intensive surveys* Ithaca Cornell Univ Press 1947 pp 80 310
 - 8 ——— and Kahn R L *The people elect a president* Ann Arbor Survey Research Center Univ of Michigan 1952
 - 9 ——— and Katona G A national survey of wartime savings *Publ Opin Quart*, 1946 10, No 3 373 381
 - 10 ——— and Metzner C *Public use of the library and other sources of information* Ann Arbor Institute for Social Research, Univ of Michigan 1950
 - 11 Cantril H Public opinion in flux *The Annals*, 1942, 220, 136 152
 - 12 Cartwright D Some principles of mass persuasion Selected findings of research in the sale of United States War Bonds *Hum Relat* 1949 2 No 3 253 267
 - 13 ——— *The selling of government bonds to the public* Unpublished manuscript 1950
 - 14 Dodd S C A measured wave of interracial tension *Social Forces*, 29, 3 1951 281 289

- 15 ——— The Washington Public Opinion Laboratory *Publ Opin Quart*, 1948, 12, 118 124
- 16 Fisher B R, Metzner, C A and Darisky B *Public response to peace time uses of atomic energy. Vol 1 Community differences* Ann Arbor Survey Research Center, Univ of Michigan 1951
- 17 Fisher, B R, and Withey, S B *Big business as the people see it* Ann Arbor Survey Research Center, Univ of Michigan, 1951
- 18 Goldsmith, S F Appraisal of basic data available for constructing income size distributions *Conference on research in income and wealth Vol 13 Studies in income and wealth* New York National Bureau of Econ Res 1951, pp 267 377
- 19 Hauser, P M, and Leonard W R (ed) *Government statistics for business use* New York Wiley, 1947
- 20 Hyman, H Do they tell the truth? *Publ Opin Quart*, 1944-1945 8, 557 559
- 21 Katona, G Financial surveys among consumers *Hum Relat* 1949 2, No 1, 3 11
- 22 ——— Expectations and decisions in economic behavior In Lerner D, and Lasswell, H D (ed) *The policy sciences* Stanford Stanford Univ Press 1951, pp 219 232
- 23 ——— *Psychological analysis of economic behavior* New York McGraw Hill, 1951
- 24 ———, and Fisher J A Postwar changes in the income of identical consumer units *Conference on research in income and wealth Vol 13 Studies in income and wealth* New York National Bureau of Econ Res 1951, pp 62 122
- 25 ———, and Morgan J N Quantitative study of factors affecting business decisions *Quart J of Econ*, 1952 66 67 90
- 26 Katz D, and Hyman H Morale in war industry In Newcomb T M, and Hurlley, E L (ed) *Readings in social psychology* New York Holt 1947, pp 437 447
- 27 Kinsey, A C, Pomeroy W B and Martin C E *Sexual behavior in the human male* Philadelphia Saunders 1918
- 28 Klein, L R Estimating patterns of savings behavior from sample survey data *Econometrica*, 1951 19 438 451
- 29 Lansing J B, and Withey S B Analysis of consumer demand from repeated interviews and reinterviews To appear in *Studies in Income and Wealth*, National Bureau of Economic Research

BIBLIOGRAPHY

-
- 1 Angell R C The moral integration of American cities *Amer J Sociol* 1951 47 No 1 Part 2 1 140
 - 2 Belknap G and Campbell A Political party identification and attitudes toward foreign policy *Publ Opin Quart* 1951 52 15 No 4 601 623
 - 3 Cahalan D *Validity of behavior reports in opinion surveys* A paper presented at Amer Stat Assoc Meetings December 1949
 - 4 Campbell A The uses of interview surveys in federal administration *J Soc Issues* 1946 2 No 2 14 22
 - 5 ——— Factors associated with attitudes toward Jews In Newcomb T M and Hartley E L (ed) *Readings in social psychology* New York Holt 1947 pp 518 527
 - 6 ——— *Attitude stability and change a re-interview study of the national population* A paper presented at Amer Psychol Assoc Meetings September 1948
 - 7 ——— Eberhart S and Woodward P *Public reactions to the atomic bomb and world affairs Part II Findings of the intensive surveys* Ithaca Cornell Univ Press 1947 pp 80 310
 - 8 ——— and Kahn R L *The people elect a president* Ann Arbor Survey Research Center Univ of Michigan 1952
 - 9 ——— and Katona G A national survey of wartime savings *Publ Opin Quart* 1946 10 No 3 373 381
 - 10 ——— and Metzner C *Public use of the library and other sources of information* Ann Arbor Institute for Social Research Univ of Michigan 1950
 - 11 Cantril H Public opinion in flux *The Annals* 1942 220 136 152
 - 12 Cartwright D Some principles of mass persuasion Selected findings of research in the sale of United States War Bonds *Hum Relat* 1949 2 No 3 253 267
 - 13 ——— *The selling of government bonds to the public* Unpublished manuscript 1950
 - 14 Dodd S C A measured wave of interracial tension *Social Forces* 29 3 1951 281 289

54 Research Settings

- 30 Lazarsfeld P F The use of panels in social research *Proc of the Amer Philos Soc* 1948 97 405-410
- 31 ——— Berelson B and Gaudet H *The people's choice* New York: Columbia Univ Press 1948
- 32 Likert R The sample interview survey In Dennis W (ed) *Readings in general psychology* New York: Prentice Hall 1949 pp 427-434
- 33 Logan W P D Illness incapacity and medical attention among adults 1947-1949 *Lancet* 1950 258 773-776
- 34 Maccoby E E Interviewing problems in financial surveys *Int J Opin and Attitude Res* 1947 1 No 4 31-39
- 35 Moss L *The social survey and public administration* A paper presented to the World Association for Public Opinion Research September 1951
- 36 School of Public Health Department of Public Health Statistics *An investigation of the causes of home accidents* Ann Arbor: Univ of Michigan March 1952
- 37 Schweiger I The contribution of consumer anticipations in forecasting consumer demand In *Studies in income and wealth* to be published by National Bureau of Economic Research
- 38 Skott H E Attitude research in the Department of Agriculture *Publ Opin Quart* 1943 7 280-292
- 39 Slater P *The social survey survey of sickness, October 1943 to December 1945* London: Central Office of Information 1946
- 40 Star S A and Hughes H M Report on an educational campaign: The Cincinnati Plan for the United Nations *Amer J Sociol* 1950 55 389-400
- 41 Stouffer S A Guttman L Suchman E A Lazarsfeld P F Star S A and Clausen J A *Measurement and prediction Studies in social psychology in World War II 4* Princeton: Princeton Univ Press 1950
- 42 ——— Suchman E A DeVinney L C Star S A and Williams R M Jr *The American soldier adjustment during army life Studies in social psychology during World War II 1* Princeton: Princeton Univ Press 1949
- 43 Survey of Consumer Finances *Federal Reserve Bull* 1946-1953 Vols 32-39
- 44 Survey Research Center *A national survey of the informed public* Ann Arbor: Univ of Michigan 1948
- 45 ——— Methods of the survey of consumer finances *Federal Reserve Bull* 1950 36 2 795-809

- 46 ——— *Industrial mobility in Michigan* Ann Arbor Univ of Michigan, May 1951
- 47 ——— *Relevant considerations in recent home purchases* To be published by the Housing and Home Finance Agency
- 48 U S Bureau of Agricultural Economics, Program Surveys Division *The victory tax and payroll deduction of bonds* January 1943
- 49 Whelpton, P K, and Kiser, C V Social and psychological factors affecting fertility, III The completeness and accuracy of the Household Survey of Indianapolis *Milbank Memorial Fund Quart*, 1945, 23, No 3, 254-296
- 50 Withey, S B *Consistency of immediate and delayed report of financial data* Unpublished Ph D thesis, Univ of Michigan 1952

Field Studies

Daniel Katz

Field studies are similar to nation wide surveys in that they have opened up new possibilities for the development of social psychology and the social sciences. On the one hand, they are breaking down the narrow walls of the traditional experimental laboratory in the application of a research approach to complex problems of human relationships. The effect is twofold (1) our scientific knowledge is increasing as a result of the direct study of field situations and (2) the psychological laboratory is beginning to include in its experimentation social and group variables.

On the other hand, the potentialities of surveys and field studies for the nonlaboratory social sciences are even greater. These disciplines have long dealt with complex and significant social problems but they have had to rely either on uncontrolled observation or on data collected for practical rather than scientific purposes. Thus they have dealt with secondary sources, such as crime statistics or census materials, in which their research designs are imposed upon data already gathered. Field studies and surveys permit the introduction of controls and of research objectives into the data collection itself. This means that both the problem under investigation and the types of observations and measures to be taken can now be under the control of the social researcher. A science which can gather its own data according to its own research interests, in addition to availing itself of existing records, is at a tremendous advance.

tage over a discipline which skips this important part of the scientific process. When the practicality of the additional costs of controlled data collection in the social sciences is established, we may witness revolutionary developments in these fields.

THE RELATION OF FIELD STUDIES TO SURVEYS

Although it is not easy to draw a fine logical distinction between a survey and a study of a field situation, there are practical differences which call for somewhat different techniques and skills. The difference is roughly between the greater scope of the survey and the greater depth of the field study. More precisely, two essential distinctions can be made. In the first place, the survey always attempts to be representative of some known universe and thus attempts, both in the number of cases included and in the manner of their selection, to be adequately and faithfully representative of a larger population. This emphasis on sampling may or may not be found in a field study, which is more concerned with a thorough account of the processes under investigation than with their typicality in a larger universe. In a survey we always ask about the relative incidence, or distribution, of social variables or personality characteristics in the larger group with which we are concerned. The explanation of trends in population increase, of economic booms and depressions, of the amount of unemployment, of social change generally, must be understood in the context of the country as a whole, so sampled that the many subgroups are properly represented and the relative weighting of factors as they contribute to the total outcome, is ascertained.

A second and more important difference is that in the field investigation we attempt to study a single community or a single group in terms of its social structure—i.e., the interrelations of the parts of the structure and of the social interaction taking place. (2) The survey, to the extent that it deals with such interrelations and interaction, does so through a study of the final outcome. The on-going social and psychological processes are inferred in the survey from their statistical end effects. In the field study, however, attempts are made to observe and measure the on-going processes more

directly. Specifically, this means that the field study either attempts observations of social interaction or investigates thoroughly the reciprocal perceptions and attitudes of people playing interdependent roles. Thus a field study will provide both a more detailed and a more natural picture of the social interrelations of the group than does the survey.

Studies of attitudes toward labor management problems furnish an example of these two types of approach. A national cross section survey would report the incidence of attitudes toward labor and toward management for the nation as a whole and would seek to get some account of the distribution of these attitudes among the subgroups in the population. From such a study we would know what are the typical attitudes of workers all over the nation, of workers in unions as against workers not in unions, of farmers, of middle class groups, and of owners and industrialists. A field study concerned with the same problem might deal with a single plant and would examine the social structure of both the union and the company. One focus of the investigation might be the power structure and the patterns of influence and communication within the union; a second focus might be similar relations within the company; still a third would be the relations between the two structures. Systematic interviewing would attempt to get at the reciprocal perceptions and attitudes of workers, foremen, stewards and higher officials, and some observation might be planned of the interactions occurring in the factory between worker and foreman, worker and steward, steward and foreman, etc.

Obviously, the field study and the national survey are not so much alternative ways of studying problems as they are supplementary procedures which can be used most effectively in combination. There are two major advantages in using both methods in the same problem area. First, we know more about the degree of generality from the findings of the field study if we know how the specific situation studied fits into the national pattern. If we knew, for example, how the people of Yankeeville, studied by the Warner group, compare with the population as a whole, we could interpret the findings more wisely (22). Secondly, the national survey and the field study each produce findings for hypotheses which can be more adequately tested by use of the other approach. For example, a national survey on class structure may suggest specific motivational

processes at work in a certain social class, a drive toward power rather than toward economic security in the top economic brackets. The confirmation of this hypothesis, however, is easier to achieve by working with a subgroup of the national population. On the other hand, after definite motivational processes have been studied in detail in a subgroup, the conception of the variables may be so sharpened and their measurement so facilitated that it may now be possible to utilize measures of these variables in a national survey.

TYPES OF FIELD STUDIES

There are many types of field studies, but the method has been most widely used by the anthropologist in his study of primitive societies. The sociologist, influenced by this type of natural observation, has made detailed studies of parts of his own society and has often been able to add some degree of measurement to the more interpretative anthropological approach. Finally, the social psychologist has emphasized the importance of quantification and verification of observation even in studies conducted outside the laboratory. The most important dimension, then, on which field studies can vary is the degree of measurement they represent, ranging from the extreme of the interpretative anthropological description of a primitive society to an investigation employing standardized quantification of data collection in the form of observational scales for recording behavior and attitude scales for the measurement of beliefs and feelings.

An illustration of an anthropological field study which attempts a functional analysis rather than a sheer descriptive account is to be found in B. Malinowski's investigations of the Trobriand Islanders (16). Malinowski lived among this Melanesian people, observed their activities at first hand and spent many hours talking to a number of native informants. He presents both a detailed description and a sociological explanation of their economic activities, their social organization, their myths and ideologies, and their psychological character structure. He reports, for example, the central role of reciprocal obligations in their economic, legal and ceremonial life.

The islanders, living near the shore, will have partners in the

inland villages so that there is constant interchange of fish and fresh vegetables carried out with due ceremonial and public display. Similarly in the fishing village there is a mutual set of obligations with respect to the fishing canoe manned by a number of natives under the ownership of one man but with effective rights on the part of the crew with respect to time of fishing and a fair share in the catch and with duties toward the maintenance and care of the vessel. Apparently the binding force of the social norms of the community comes not from legal penalties or from sheer conformity to custom but rather from a sustained give and take with both parties to an arrangement benefiting thereby and from the reinforcement of this relationship through ceremony and public display. Thus a man who can present his partner with a lavish heap of food both ensures a good return in the future and is highly regarded by his fellows for his prowess and generosity. Malinowski concludes that reciprocity is the all important principle in the social norms of this primitive society.

Interesting and plausible as Malinowski's interpretation is it lacks the finality of scientific generalization even for Melanesian society. The hypothesis about the importance of reciprocity for the maintenance of social norms needs to be tested in systematic fashion in the Melanesian culture in relation to other social processes which may be functioning. For such a systematic test it is necessary to have some measurement of these processes. It should be noted however, that Malinowski's functional analysis has definite advantages over many descriptive accounts of cultures. This type of analysis utilizes theoretical concepts which point fairly directly to observable social interactions. Hence the interpretation is readily converted into testable hypotheses.

A sociological application of the anthropological approach is found in the Lynd's study of Middletown which reconstructs the life of an American community on the basis of the intensive investigations of a small team of field workers (15). The field workers lived in the community and participated as fully in its life as they could employing participant observation as one of their major methods. In this process informal interviewing occurred frequently. In addition they conducted a thorough examination of all documentary materials including school records, records, court files, organizations were read.

MLSU - CENTRAL LIBRARY



17810CL

not only for the current period of the study but for earlier years as well. Similarly, past and current files of newspapers were consulted, and histories of the state, county, and city were studied. Considerable emphasis was placed upon understanding the community in terms of its history. Diaries were examined for the early years to supplement the other historical records. Moreover, where no statistical materials were available, the field staff compiled data on wages, steadiness of employment, club membership, church attendance and membership, attendance at motion pictures, etc. This introduction of quantitative materials was made more rigorous by carefully planned interviews with a sample of working class wives and a sample of businessmen's wives. Written questionnaires were also sent to more than 400 clubs and to three fourths of the senior high school population.

From this mass of data and source material, the Lynds derived their account of the major activities of the community, the trends in its development, the pattern of conflicts in its life. They found a differential rate of change in the performance of various basic activities, with the greatest changes in the economic pursuits of getting a living. The general pattern of change was from the business class to the working class, with the working class often showing today the habits of the business class of a generation ago. There are instances, however, when this process is reversed and practices of the lower income groups have been taken over by the upper income groups. In general, the currents of change seem erratic, with respect both to direction and to the differential rates for different aspects of life. The difficulty of maintaining some sort of equilibrium under these conditions of stress emerges as a major problem. One solution which does appear, though it is not systematically utilized, is a sidling procedure in which an innovation first appears as an optional alternate mode of adjustment and then gradually replaces the older mode.

The Lynds have gone beyond the traditional anthropological method in utilizing quantitative techniques to supplement their qualitative materials. Their statements about the activities of the community and the characteristics of its people are often derived from statistical tables. Their conclusions and their more interpretive picture, however, are drawn heavily from a consideration of the qualitative information and from their own experiences as

participant members of the community. Thus not all of their conclusions can be readily confirmed by other investigators because of the difficulty of duplicating the exact procedures on which the conclusions are based. In general however, in its attempt to document its observations with facts and figures this study represents a real advance toward more objective scientific methods. What emerges very clearly in this attempt is the distinction between findings or data and interpretations which are derived from the observations and data.

The pattern of the Middletown study in interweaving quantitative data obtained from interviews and questionnaires with material from secondary sources and general observations and information about the cultural setting has been followed in many sociological studies. These have often dealt with a more limited frame of reference than the total community and have contributed to our knowledge of social stratification as in the Warner studies (23) the Hollingshead investigation of adolescence and class membership in a midwestern community (9) and the Jones study of the socioeconomic basis of class in Akron, Ohio (11). An interesting combination of participant observation and systematic interviewing is found in Child's study of second generation males of Italian origin (3).

An early example of the approach of social psychologists employing more measurement but circumscribing more narrowly the field to be studied is Schanck's study of Elm Hollow (19). Schanck followed the anthropological tradition in living in the community for a period of years and becoming thoroughly conversant with its ways through observation and through long talks with informants. In addition however he used systematic interviewing in which every resident of the community was questioned in an informal interview. The views of each resident were ascertained on the same standard issues. Thus Schanck was able to quantify his findings and to state in more precise fashion the degree of relationships found. It is interesting that this detailed and quantitative approach gave an account of the community which distinguished sharply between the formal patterns of belief and behavior in institutional settings and those in more private and informal settings. Just as the Mayo (18) group found that on close inspection the formal patterns of a factory are often contradicted by the informal patterns of inter

personal relations, so Schanck found that the fundamentalistic beliefs and practices of the older religion were largely for public purposes. In private, individuals maintained a much more liberal set of beliefs and attitudes. Part of the conformity on public occasions was a matter of pluralistic ignorance, the erroneous belief held by many that the rest of the community felt differently from the way the respondent himself felt. Also important was the acceptance by the villagers of the leadership roles of the town minister and the chief contributor to the church funds. Later study showed that without the reinforcing effect of these leaders, the pluralistic ignorance in the small community was quickly dispelled, with genuine reversals in accepted patterns of behavior concerned with religious taboos.

This early community study which emphasized the individual as a unit of measurement was restricted in scope and in technical thoroughness. A social psychological field study employing more detailed measurements is Newcomb's research on Bennington College, a self contained college community (17). Measures of objective role in the college were obtained from ratings of a cross section of student judges who selected the most extreme individuals in each class on each of 28 characteristics related to community citizenship. Subjective role was measured by the individual's own view of her relationship to the community, including her awareness of differences between self and others. Individual prestige was ascertained through nomination of students, by their fellows, as most worthy to represent the college at an intercollegiate gathering. Finally, a series of attitude scales was administered to all students and was repeated for some of the students during their college careers. The attitude scales dealt with public affairs which were central to the values of the community. The results showed that, as students assimilated to the college community, they took on the values of the group. The acknowledged leaders showed greater effects of such group membership than the followers. Thus there was good correspondence between the attitudes of students valued by the community and objective role played by the student. Moreover, subjective role indicated that within the objective pattern there were characteristic personality modes of adjustment which were the resultant both of the present situation and of past methods of relating to one's fellows.

These four studies indicate that the conflict between the anthropological and the quantitative approach can be resolved, though the differences in methods should not be glossed over. The anthropologist or sociologist who has familiarized himself thoroughly with a culture or community, who has lived in it, observed its people talked with them at great length, studied its history, and immersed himself in all available materials can give a picture of the functioning group as a whole. He can make insightful interpretations of its social processes. And this type of study provides a great deal of information about a community or a culture with a remarkable economy of effort. For example, Harbison and Dubin, in their case studies of union management relations, produced a remarkably informative picture of the significant variables and their interrelations (8). Similarly, Dollard (6) and Davis, Gardner, and Gardner (5) demonstrated the economical advantages of the anthropological method in their studies of caste and class. A measurement approach would involve many more field workers and many times the time and energy and would still not be able to achieve the same high level of understanding and interpretation. Moreover, the quantitative approach, because it seeks for easily measured variables, may focus on microscopic and trivial factors and miss the significant processes in group functioning.

On the other hand, the anthropological procedure represents only the first step in science, because its rich interpretations are not based on relations which have been quantitatively established. They are inferences which either represent a wholistic type of judgment or are based upon what the investigator regards as his most central observations. There is little attempt at specification of the types of data which are necessary for the measurement of a given variable. Hence, it frequently makes difficult and often impossible the verification of relations by another investigator. The history of social psychology illustrates the importance of the replication of findings in that many of its initial results have not been confirmed by later investigations. Only when we attain the level of standardizing our specifications for data can we see the extent to which reported findings are true generalizations.

This dilemma posed by the two approaches is more of a historical accident than a logical necessity. In many instances it can

be readily resolved by utilizing the anthropological approach as the initial stage in a field study. This stage can utilize to the full the advantages of seeing the situation as a whole and of attempting to grasp the fundamental relationships. From this study can come the insights which can furnish the hypotheses for later, more detailed, quantitative study. In fact, in the Festinger-Schachter-Bach study of a housing community precisely this procedure was employed (7). In the first stages, informants and informal participant observation were used, as well as observations of important group meetings. Then, in the later stages of the study, systematic interviewing of all housewives was carried out, and sociometric techniques were used to discover communication and preference patterns.

STEPS IN THE CONDUCT OF A FIELD STUDY

It is important, then, to turn to the steps in the conduct of a field study. The following model cannot be fully realized in every study. Moreover, specific studies often dictate their own procedures. But there is some advantage in breaking down an investigation into its major processes. Thus, the following phases can be examined for their relevance to a contemplated study: (1) preliminary planning, (2) the scouting expedition, or the anthropological short cut, (3) the formulation of the research design, (4) the pretesting of research instruments and procedures, (5) the full scale field operation and (6) the analysis of materials.

Preliminary Planning

Ideally, the field study should start with a period of research planning in which some tentative decisions are made about the scope of the study, its general objectives, and the timetable of its stages. As a general rule, exact formulation of research design is left to a later stage, when the results of the scouting expedition are available. Often one purpose of the field study is the obtaining of a better knowledge of the significant variables rather than the final testing of a well formulated theory. Even where the field study is a follow up of other research, however, it is important not to

freeze the design before the scouting expedition. It is difficult and sometimes impossible to know what measures are feasible in a given field setting without a firsthand exploration of the situation.

It is well to be alert to the general temptation to envisage the study too broadly and to make an unrealistic appraisal of what can be accomplished within the time and budgetary limits of the project. The scouting stage can be more valuable if there is some major focus and some restriction of area. Another temptation against which planning offers protection is the tendency to accept a given community or group for study because of its easy accessibility and because of assured cooperation from a few key people in it. These are important considerations but they should not outweigh the research objectives. It may be that the most easily accessible community is not the best place for studying the phenomena in which we are interested. If the suitability of the community for the purposes of the study cannot be decided in early planning phases it can be posed as one question for the scouting stage.

There is no agreement in practice about the use of previous research in the planning phase. The current tendency is to ignore what has been done in the past, because the researcher does not want to become contaminated by old concepts or because he regards previous studies as irrelevant and useless or because he prefers to use his time in his own research rather than in the library. Undoubtedly, in the pioneering stage of any discipline, where sophisticated methodology is relatively new, there is much justification for this point of view of moving ahead and disregarding what has gone before. But increasingly we shall want to build a science, and this can be done better if each investigator does not start anew with his own terminology and insulate himself from what others have done or are doing. This is true with respect to both substantive research findings and methodological results.

It is better to start with some general plan concerning research objectives, personnel and timing than with an unstructured program. The plan, however, should allow for changes in decisions as a result of scouting and pretesting. The fact that there is a plan, however, enables the people doing the scouting to bring back information on the types of questions about which final decisions must be made. It may not be desirable to decide at the outset which specific subgroups within a community are to be studied most

intensively and which specific processes are most fruitful for the study. But if some of these questions can be posed in advance of the exploratory scouting, much time can be saved.

The Scouting Expedition

The scouting phase of any study is the period of informal and relatively free investigation in which the field workers try to get as thorough an understanding of the important forces in the situation as they can. During this period they either live in the group to be studied or make frequent trips to observe it at first hand. The scouting expedition is, thus, not a pretest in which already formulated instruments are given a field trial. It is essentially exploratory, with the objective of finding out what the significant variables in the situation are likely to be and what types of instruments may have to be constructed to obtain measures necessary in the final study.

The advantages of employing more than one investigator for the scouting study are obvious. Not only is the single investigator limited by time pressures but his own biases need to be checked. Moreover, with a team of field workers supplementary and complementary skills can be utilized.

Although freedom for the investigator to follow interesting leads and to utilize his own ingenuity in obtaining information is the very essence of the scouting stage of a study, this is not freedom in an absolute sense of random or aimless activity. Enough is known about social groups in general so that we have some knowledge of the types of things to look for in most social situations. For example, even though the Lynds (15) envisaged their study of Middletown as exploratory, with no attempt to prove or disapprove any set of hypotheses, they did assume that there were broad categories of basic social behavior that should be observed—getting a living, making a home, training the young, using leisure in various forms of play and art, engaging in religious practices, and engaging in community activities. As a rule, the broad framework of types of behavior to be studied will vary in relation to the purpose of the investigation. Nevertheless, there is much to be said for having some systematic plan in mind for scouting purposes to ensure the type of coverage of information which the nature of the study may require.

The following framework is suggested as a possible guide for

obtaining broad coverage on important aspects of group functioning. It will not, of course, have equal value for all types of field studies, but it does reflect many of the types of variables which social psychologists and sociologists are finding significant in the understanding of the specifics of group behavior.

- 1 A description of the total structure under study with respect to the major groups and subgroupings
- 2 The central value systems and goals of the total system and its various groups
- 3 The nature and types of conflicts and points of tension both with respect to the total structure and with respect to a single group
- 4 The formal and informal structure and the way in which they are interrelated
- 5 The accepted pathways to group goals including
 - a the logical relation between paths and goals
 - b the remoteness of paths from ultimate goals or the number of subgoals between a group's activity and its ultimate goal
 - c the degree of fixation upon one or two main paths and the range of permissible alternate routes
- 6 The degree of autonomy of functioning of the parts within the total structure and the nature of their dependency upon one another and upon the larger whole
- 7 The nature of the dependency of the structure under study on the society or larger unit of which it is a part
- 8 The power or influence patterns within the structure and its subgroups
- 9 The nature of the group sanctions and the degree and basis of their acceptance by group members
- 10 The patterns and channels of communication within the structure and the substructures

It is also helpful for the field worker to be trained in informal ways of gathering information. There are a number of practical procedures which can be followed, though their usefulness will vary from study to study. The following procedures should be kept in mind by field workers when they go into a community, a group, or

an industrial plant to carry out the anthropological part of the investigation

1. Contacts should not be limited to a narrow segment of informants. People are very limited in their information by their daily social roles. They not only lack knowledge of the activities of others but they are biased by the positions they occupy. Hence it is important to be in communication with some person from every important subgrouping and every important type of social role. This has been the classic error in diplomatic intelligence in the past; observers in a legation post have moved among people like themselves, representing the upper social strata of the country they are studying. In general, when we enter a new community we tend to seek out people very much like ourselves. So, too, does the inexperienced field worker. Even interviewers with specified quotas of respondents to obtain will, if not checked by certain controls, bring back an undue number of interviews with people of the same religious and socioeconomic characteristics as themselves. Hence, the field worker should be alert to the problem of obtaining a wide coverage of informants. Opinion studies which have not met the requirements of precise sampling have often been surprisingly accurate because they have obtained a fairly wide spread of respondents from all important types of groupings.

2. Informants who themselves have a wide range of contacts should be utilized. The person who by virtue of his role or his personality has a high rate of contact may have especial usefulness for the field worker. People involved in communication activities maintain many contacts and often have information which does not appear in their official reports. Newspaper reporters, for example, can readily describe the pattern of leadership in the community and identify the hierarchy of informal political bosses.

3. Informal leaders, as well as the people in positions of formal leadership, should be located and consulted. The account of the formal leader always needs to be checked and supplemented with what the informal leader can contribute. Not only is the informal leader often in possession of facts and interpretations not known to the official, but he may also be in a better position to express freely what he does know. For example, in one field study of union management relations one of the men at headquarters, holding

no official position could describe union policy very frankly. He admitted that there was some fat in the present group piece rate that the company was trying to change on the occasion of the introduction of a new assembly line. The real issue was not the attempted change but the basic intent of the company in wanting change. If the company wanted minor and reasonable modifications, that was one thing, but if this meant a new policy of pushing the union to the wall, then the men would fight against the most minute change with all the resources they could muster. And until they could get a better line on basic company intent they intended to do some experimental skirmishing with the company on the issue of new standards. The responsible union officials would not express this point of view so explicitly, but later events showed that it did represent union sentiment and union policy.

Locating the informal leaders is generally not too difficult if the field worker can spend enough time in the community under study. They are usually known to those who themselves have had some relation to the practical functioning of the groups in question. Those who have attempted to organize some community function soon discover who the key people are whose cooperation is essential. Often the officials of a rival group can readily identify the informal leaders on the other side. A plant manager who wants to introduce a new machine often knows that he must convince not only the union leaders but also the informal leader among the old time workers. And the rank and file can generally say to whom they turn for advice and direction.

4 Discrepancies in the accounts of various informants should be used as the basis for further exploration. There should be discrepancies in the information the field worker is obtaining. If all his informants give him the same story of complex group relations and functions, he is probably not covering a wide enough representation of people in different roles and different positions. The contradictions he does find, however, should determine the direction of further inquiries. He can, by questioning additional informants, find out whether the differences in the report he has obtained are a function of idiosyncratic perception and experience or a reflection of group membership and role differences. Moreover, apparent contradictions can be resolved by determining the frame of reference of the respondents who disagree.

5 Information from respondents should be assessed in relation to their social role and position, their group memberships, and their personal activities. Hence, it is important to get as much information about the informant's priority of group allegiances, his position in the power structure of the groups to which he belongs, his major roles, as well as his own aspirations and goals in life. One principle that generally holds true for hierarchical structures is that the people at various levels in the hierarchy are generally more sensitive about the actions and feelings of those immediately above them than of those below them. Advancement up the ladder depends upon an upward orientation and an ability to relate oneself effectively to one's superiors. Hence, a foreman in a plant may know more about the way of thinking of his immediate bosses than of his own men.

In addition to motivational biases, which need to be known before information can be assessed, is the factor of the amount of knowledge which the informant can be expected to possess on the basis of his contacts and experiences. This was the reason for the previous emphasis upon finding respondents who are very active and who enjoy wide contacts in the community. One caution that needs to be stressed has to do with the *positional lag in information*. The man who has moved up in a hierarchical structure knows through his own experience and through the important contacts he had in the local group, about the lower levels from which he has come. In his new role, however, he has generally lost these points of communication. Nevertheless, he often feels he can talk accurately about his former level of association. Similarly, the general in the army will talk authoritatively about the problems of his men because twenty years ago as a lieutenant he was close to them and understood their way of thinking. This informational lag is less true of organizations based upon functional representation, where the leader must report back to the group which elects him. This procedure immediately forces contact between the leader and the people below him. But if the organization lacks functional representation, the information which high level leaders may give about the lower levels may be irrelevant and inaccurate.

6 Ideally, it is desirable to spend considerable time in participant observation. Reports of informants and information derived from secondary sources need to be supplemented with living in the

community participating in its activities and constantly observing what people actually do in specific situations. There are frequently practical difficulties with participant observation, since it adds greatly to the length of time the scouting phase will take. But there is no good substitute for having field workers actually live in the community and perform some of the roles they are interested in studying. Empathic understanding of the problems faced by the people under study cannot be obtained fully through hearing about experiences from others or even from direct observation. Thus, if we are making a study of an industrial plant, the field workers will have a much better account of the situation if they can actually spend some time both in the manager's office and on the production line. In the absence of experiences from participation, the field worker should spend as much time as possible in direct observation. He should attend meetings of organizations and observe people in group situations. There are interesting discrepancies between what people say in isolation and the way in which they behave when they are under group pressure.

7. Personalized and private beliefs should be sought as well as the socially accepted climate of opinion. In an attempt to be helpful and objective, informants will report the accepted point of view about which there is public agreement. This public agreement may represent what people are supposed to believe and say: the world of newsprint, the publicized and semiofficial version of the state of affairs. Now it is essential to know this public climate of opinion since it does affect social behavior. But it is also important to get below this first level to the more private beliefs and actions of individuals. The field investigator should attempt to get from respondents their own private views and their own personal behavior as well as the accepted climate of opinion. And behavioral observation procedures can be helpful here.

In the 1948 presidential election almost everyone except President Truman had accepted the inevitability of the election of Thomas E. Dewey. The press, the periodicals, people in official position, even professional politicians had accepted this public fiction. Although the Gallup poll showed Dewey ahead by only 5 percentage points, with 12 percent of the people undecided, the pollsters themselves were victims of the fiction. So powerful did this myth grow through constant social reinforcement that the conserv

ative betting fraternity was giving odds of 55 to 1 that Dewey would be elected

The field worker needs to be aware of this type of public fiction. It can exist where a belief has become so prevalent that few would dare to challenge it. It can also exist in the area of taboo subjects, such as sex. The Kinsey report, although not based upon representative sampling procedures, does raise the question of the extent to which accepted beliefs about sex practices are public fictions.

8 Full records should be kept by field workers. Part of the discipline of the investigator is rigorous note taking and the setting aside of at least two periods daily during which the notes are elaborated into a full report. No matter how excellent his memory, the worker cannot reconstruct from his notes his original observations after a lapse of time without some losses in completeness and accuracy. This is especially true in the field situation where a constant succession of similar experiences may maximize retroactive inhibition and inaccuracies.

9 Initial impressions and global judgments should not be discarded. Although detailed documentation is the goal even during the scouting period, it is nonetheless true that this is also the period when maximum play should be given to overall impressions. As the Gestalt psychologists have demonstrated so effectively, the human mind does grasp things as a whole. But this type of wholistic perception tends to be neglected in our scientific efforts at precision. In the scouting stage, however, field workers should be encouraged to record their initial impressions. These first judgments can be surprisingly useful because the situation can sometimes be perceived in its main outlines at the very start. As the exploratory work progresses there is a tendency for the details to obtrude themselves. Therefore, there is some point for the investigators to try a summing up at stated intervals to make them see the whole picture again.

10 Available records and secondary sources should be studied carefully, and the operational procedures for deriving such records should be examined. Not only are such existing materials of great value in the understanding of the situation but they sometimes can be used as measures of variables in the larger study. For example a field study of an industrial situation may well want to inquire fully into productivity and other records maintained by the company being studied. When the report is made of the scouting expedi-

tion, it is not enough to know that productivity records exist on individual workers. It is essential to know what these records are based upon, to what degree the productivity of the worker is set by the pace of the machine or the assembly line and to what degree by his own efforts, how comparable productivity records are for individuals performing different tasks, how stable are the productivity differences reported over time, etc. Unless the operational meaning of these measures is known, it is impossible to construct a research design which will utilize such records.

The Formulation of the Research Design

As the results of the scouting exploration become available, the design of the final study can be worked out more exactly. There are advantages in developing the design as the scouting proceeds rather than making it a separate step in a temporal sequence. This permits of some interaction between the possible theoretical objectives and the realities of the field situation. At some point, of course, final decisions must be made about research objectives and procedures for the full scale study, and such decisions call for a thorough consideration of all the findings from the scouting expedition.

Roughly speaking, studies are of two major types: exploratory and hypothesis testing. The exploratory study attempts to see what is there rather than to predict the relationships that will be found. It represents the earlier stage of a science. From its findings may come knowledge about important relationships between variables, but the more definite proof of these relationships comes from hypothesis testing.

For example, in a field study of industrial morale we may be interested in the factors related to productivity. If the study were of an exploratory type, it would not start with clearly defined notions about the relationships to be found. It would set a broad net and include measures of a wide variety of perceptual and motivational factors in the hope that some of these measures would show a relationship to productivity. If the study were of the second type—namely, hypothesis testing—we would start with a well formulated notion that under specified conditions productivity would vary directly with a given factor or factors—perhaps the group standards of the face to face members of a work section plus their involve-

ment in the group. In this second type of study, we would develop detailed measures of these independent variables and would make exact predictions for the productivity of work groups varying in group standards and solidarity. We would also specify the conditions which have to be held constant for these predictions to be realized. Since these conditions may not be held constant directly, we would measure them to achieve some statistical control over their effects.

Ideally, the testing of hypotheses is more suited to laboratory experimentation, and exploratory discovery to field studies and surveys. This does not mean, however, that field studies should confine themselves wholly to exploratory procedures. The scouting stage can often be used as the more purely exploratory part of the investigation, and some degree of hypothesis testing can be employed in the larger operation to follow. Moreover, there are occasions when the field approach can be used for very important hypothesis testing, as in the "natural experiment" (see Chap. 3). But it is nonetheless true that the great strength of the field type of study is its inductive procedure, its potentiality for discovering significant variables and basic relations that would never be found if we were confined to research dictated by a hypothetical deductive model. Thus, the field study and the survey are the great protection in social science against the sterility and triviality of premature model building.

It is possible, of course, to combine both exploration and hypothesis testing in a single field study. One major set of hypotheses can be investigated at the same time that other materials are gathered for exploratory purposes. This has the advantage of protecting the study from failure if inconclusive results are found with respect to the hypotheses. The exploratory materials then become the *safety factor*. The disadvantage of this compromise is that it attempts to combine two studies in one investigation, sometimes to the detriment of both.

Even an exploratory study should be so designed as to provide as definite information as possible for a set of research objectives. There are at least two levels of exploratory studies. At the first level is the discovery of the significant variables in the situation; at the second, the discovery of relationships between variables. Even at the first level it is important to delimit the area to be studied and to introduce controls into the data collection process. Exploratory studies which do not set limits for themselves have limits im-

posed by various practical matters, some of which are not realized by the investigators

For the exploratory study aimed at the discovery of variables rather than relationships factor analysis is often urged as the best method of finding out the unitary and independent factors in the situation. From a design point of view, much more is known about the assumptions of factor analysis on the side of statistical treatment of materials than about the assumptions on the data collection side of the process. The tendency in the applications of factor analysis to social settings is to throw all sorts of measures of various degrees of precision and validity into the hopper of factor analysis and to depend upon statistical sophistication to grind out meaningful entities. Controls on the data collection side, such as measures taken under standardized conditions and an adequate sampling of situations are disregarded. Although factor analysis is a powerful tool for handling statistical materials, it is of very limited use in field studies unless the measures we employ in the first instance are defensible. The major need, then, in the design of the exploratory field study is the provision for controls in the observation of behavior and in the recording of respondents' ideas, perceptions, attitudes, sociometric choices, etc.

These controls should be concerned with standardized or comparable conditions under which observations are made and interviews taken and with measures of reliability for the data gathered (see Chapter 6). This implies a fair degree of specification of the cues the investigator is to use in coding the behavior he observes and a certain degree of structure in the interviewing situation. The freedom of the scouting phase is over, and we now need measurements of the factors we are describing as important in the area of our study. Another major requirement in this first level type of exploratory study is adequate representation of the relevant variables in the situation. This calls for thorough and even coverage of the many aspects of behavior which seem to be related to the main problem under investigation. Only under these conditions will factor analysis or some similar technique have a real opportunity of discovering the significant unitary factors.

In the second type of exploratory study, where the objective is the discovery of relationships, there is less concern with adequacy of coverage of behavior and less interest in the use of factor analysis.

Thus it resembles hypothesis testing in resting its case upon the relationships discovered rather than upon the precise use of mathematical techniques. The major difference between such an exploratory study and the hypothesis testing investigation is that in the former there are no specific predictions of relationships based upon theoretical derivations. The researchers do have hypotheses in mind, but these are not precisely formulated. In a study of class structure in a community, for example, we may start with the general assumption that a significant motivating factor in class identification stems from the economic role which the individual plays. But we may not be prepared to specify what we mean by economic role, or what other roles may account equally well for psychological class identification. Therefore we plan our research so as to study the many possible types of economic role, including the part the individual plays in consumption, in the technical aspects of production, in the social aspects of production, etc. Within the broad frame set by our research objective, we hope to find some significant relationships. Or, in a study of industrial morale, we may be concerned with the in-plant factors which are related to worker satisfaction. We shall include all the important aspects related to the job and the plant from wages and working conditions to type of immediate supervision and congeniality of fellow workers. Then, in analysis we hope to find significant relationships between worker satisfaction and some of these in-plant factors.

In this sort of exploratory study, the design should be so constructed that measures are available for all relevant dimensions of the area under investigation but the study should be confined to a limited type of problem. It may be that a whole set of variables, which have been omitted as not belonging to the area under investigation, have more to do with the dependent variable than the factors which are studied. For example, economic role conceivably may not be as important in class identification as sociometric personal preferences or the length of residence of the family in the community, or the number of ancestors who fought in the Revolutionary War. But it is a mistake to believe that one study is going to be able to account for all the variance in complex social phenomena. It is much more effective to take one central set of variables and investigate them as thoroughly as possible than to try to study the universe in one piece of research. This is a widely accepted principle among

research workers save when they evaluate the research of other people. Then objections are often raised to studies because they omit many significant causal determinants. Thus for example to the extent that the California studies of prejudice (1) are concerned with the relation of personality dynamics to discriminatory attitudes it is not legitimate to criticize this objective because group membership, economic interest, etc. are not investigated. We make progress in science not by trying to solve all problems at once but by going at things one step at a time.

The best opportunity for the use of hypothesis testing is on the occasion of the natural experiment. The difficulty with the use of hypotheses in field studies is the inability to determine causal relationships with any definiteness, since most of our measures are not taken with respect to systematic changes in some ascertained independent variable. Now a natural experiment is a change of major importance engineered by policymakers and practitioners and not by social scientists. It is experimental from the point of view of the scientist rather than of the social engineer. But it can afford opportunities for measuring the effect of the change on the assumption that the change is so clear and drastic in nature that there is no question of identifying it as the independent variable at least at a gross level.

For example, during World War II many Japanese who were permanent residents of the Pacific Coast were uprooted from their homes and communities and assigned to war relocation camps. In his insightful book *The Governing of Men*, Alex Leighton (14) describes the effects of this uprooting in a specific camp. Although no measurements were taken, Leighton's observations showed a significant role reversal in Japanese American family structure. When the group was part of the American society, the American born Japanese had assumed the dominant position in the home and the community over the older family members who were Japanese born. A real departure from Japanese tradition. With the uprooting and rejection by the dominant culture, the leadership function reverted to the older people.

Our best knowledge of the effects of contact and personal communication on racial prejudice comes from the natural experiments of the Army during the World War II in abolishing racial segregation practices in certain combat units (20). The Research Branch

of the War Department was able to measure some of the effects of contact under these conditions of group survival, although the study was not a natural experiment in that its measures came after the event. Nonetheless, when a planned change is known, it is feasible to formulate hypotheses in advance and to take continuing measures and observations of the on going change. The full advantages of the natural experiment remain to be exploited.

It is true, of course, that to the extent that psychologists are allowed to participate in a planned change by being fully informed and by being permitted to take detailed measures, the way is open for them to influence the planning of the change. Thus we have a bridge between the natural experiment and the field experiment. An interesting illustration of the possibilities here is found in the Curle Trist study of the rehabilitation of returned British prisoners of war (4). Many of these men were having real difficulty in readjusting to their old roles in the family and community. Hence, camps were set up in which a group of men could live for a period and maintain something of the norms and roles which they had developed as prisoners of war. But contacts with civilian life were gradually established to facilitate adjustment to the civilian world. This natural experiment was studied very carefully and was to some extent carried out on the advice of psychiatrists and psychologists.

The major advantage of the natural experiment over the laboratory or planned field experiment is that the manipulation of variables is much more powerful. The real world can and does produce role reversals, drastic changes in group norms, institutional revolutions, and group conflict in a fashion impossible in the laboratory. The design difficulties in the natural experiment, however, are much greater than in the laboratory experiment. We generally lack a control group whose comparability to the experimental group is assured. Hence provision should be made in the design for obtaining measurements on as well matched a control group as possible. This does not guarantee that the experimental and control groups will be truly equated on everything but the independent variable. It does, however, increase the probabilities that the predicted results if confirmed are valid.

A second provision to make is the detailed measurement of the degree to which the independent variable is manifest in the subgroups under study. If for example, in the Army experiments we

were studying the effect of contact, we would want to know how many Negroes were actually introduced into a given combat unit. In other words, we need a measure of the amount of the independent variable which is independent of the measure of its effect. A third provision should allow for continuing observations during the course of the change process. Such observations can be helpful in the interpretation of results because they may show many processes which intervene between the initial change and the final outcome.

The fourth provision—and by far the most important in hypothesis testing in a field study—is the degree of elaborateness and specificity of the predictions which are made in advance on the basis of theoretical expectation. When we are conducting an exploratory study, we are handicapped in our *ex post facto* analysis in interpreting the relationships which do appear. The direction and meaning of such relationships can often be interpreted in many ways. But in hypothesis testing, where we have specified clearly and in detail the relations we expect to find, we have a guarantee against the inadequate controls and the loose type of measurement we may have had to employ. The guarantee applies only to clear cut positive or negative findings, not to lack of correlation. If our predictions are borne out faithfully, then the relationships discovered are not a function of spurious measures or erroneous interpretation but are in all probability a true account of causal connections. But if our predictions are neither clearly proved nor disproved, then we can say little about the lack of relationships, since they could easily result from deficiencies in method. Positive results are more convincing when the hypothesis has been elaborated into a set of interdependent propositions. Moreover, the prospects of confirmation are greater when such detailed predictions are set up to account for the differential behavior of the various subgroups and the various types of people under varying conditions. The design of the study must, of course, be tailored to obtain measurements of such factors.

An excellent check in the research-design stage in studies trying to measure relationships is the setting up of the tables at the time the design is being elaborated. Especially if the study contains some hypothesis testing, it is advisable to attempt to anticipate the tabulations and cross tabulations, or correlations, which confirmation of the hypotheses requires. By actually going through the mechanics of

setting out such tables, the investigators are bound to discover complexities of a variable which need more detailed measurement and qualifications of hypotheses in relation to special conditions. The less the study attempts to make specific predictions, the less detailed the preliminary sketch of the tables needs to be.

Regardless of the degree of hypothesis testing of the field study, the design should exploit fully three of the natural advantages of such investigations. The first advantage is that the field study tends to continue over a period of time, so that it is possible to maintain continued observation. Thus it follows that the timing of certain variables may be ascertained. We can do very little in *ex post facto* analysis, where we are dealing with variables which are untimed. If we find in industry that supervisors who follow democratic human relations skills have sections with larger productivity, we do not know with any assurance which is cause and which is effect. We can assume from general psychological knowledge that it is more likely that the skills were not produced by higher productivity, but we are much further along scientifically if we can time the occurrence of these two variables. In this case it would mean following given supervisors as they are transferred from section to section. Even where the time period of the field study is not itself very long it can afford chances to check on the timing of factors through the consultation of records and the use of the memories of a number of respondents. Unless the design clearly specifies the types of such measures for the timing of variables it will be difficult in later analysis to pin down such information.

A second advantage of the field study which should be utilized is the opportunity for the direct observation of interaction and of social relationships. We make inferences about social process and social structure from surveys but in the field study we can observe these factors more directly. If, in our study of the college community, we are interested in the effects of group membership in various organizations we should not be content with getting the attitudes of these group members in isolation. We should have observers at group meetings to record how people actually behave in the group situation.

A third advantage of the field study is the important resource of going beyond measures obtained from a single instrument. The correlations from a single measuring instrument may be influenced

by some halo in the application of the instrument. Leaders who report good communication practices with their followers may also be the leaders who report good communication practices with their own superiors. This may mean not a true generality of relating oneself to others in the organization but optimism or conventionality in answering the interviewer's questions. A field study permits the obtaining of reciprocal perceptions and interdependent reactions from different groups of people whose behavior is interrelated to make up a social structure. Agreement in perception on the part of people standing at various points in the hierarchy gives greater confidence in the validity of the report. For example, when workers, foremen, and stewards in a department of a factory all agree about whether the foreman or steward has the greater power in that department, we are on much safer ground than if we had only the foremen or the stewards reporting on the situation. This is important, not only for our knowledge about this differential power variable in itself but in its relation to other responses of foremen and workers. Moreover, discrepancies in perceptions can in themselves be meaningful psychological factors, for we can measure the perceptual distortion which people have toward some competing group or toward their leaders as it relates to feelings of hostility, lack of communication in group identification, etc. Finally, reciprocal perceptions can be put together to provide a picture of a total structure the complexity of which we might otherwise miss.

The use of independent measures in the field study should not be confined to interviews with different subgroups or types of people. It should be extended to include behavioral observation and existing objective records. Again the relationships that are found between measures obtained in these different ways are more convincing than if they all derived from a single instrument. It is not so much a question of validating interview response against behavior as it is a matter of assuring that real relationships exist between the factors that are measured. In a study of the effects of strategic bombing upon German civilian morale in the last war, it was possible to obtain a measure of exposure to bombing independent of the respondent's own report of his war experiences (12, 21). Two sources of objective information were available concerning exposure to bombing: the Air Force records of tons of bombs dropped on the town in question and the percentage of

houses destroyed in the respondent's town as ascertained by the field investigator. Thus, the interview could deal with various problems of morale during war, such as confidence in leadership, equality of sacrifice, etc., without the respondent's realizing the major objective of the interview. Hence, relationships between morale and degree of exposure to bombing could not be attributed to the halo effect of the interviewing instrument. If in designing a field study, we limit ourselves to responses from the interview for all our measures of the independent and the dependent variables, we neglect the unusual potentialities for methodological advance in the field approach.

The Pretesting of Research Instruments and Procedures

The elaboration of the research design of the study should contain the specifications for the measures required. These measures call for such instruments as interview schedules, questionnaires, behavioral scales, and forms for the gathering of information. Wherever the research objectives permit, instruments that have already been standardized in other studies should be employed. This use of common instruments would facilitate the comparison of findings from study to study. Nevertheless, it is still true that in most investigations for some time to come we shall need many new instruments developed to suit the objectives of the study.

It is essential that every new instrument be pretested before the full scale field operation. Such pretesting has three purposes: (1) to develop the procedures for applying the research instrument so that, for example, the scale or schedule can be used effectively with respect to the time it takes to administer, (2) to test the wording of questions so that they are suited to the understanding of the audience, and (3) to ensure as far as is practical, that the specific questions or observations are really getting at the variable for which a measure is needed. The pretesting of instruments is too often limited to the first two objectives since they pose readily soluble problems. There is always a genuine danger that the third objective may be slighted in favor of the interesting question: the observation easy to make, and the general ease of administering the study. The novice easily becomes discouraged by the variable that is difficult of measurement; consequently he shifts, often without

realizing the extent of his shift, from the research objective to the easily obtainable information

Another difficulty with developing measures for the variables designated in the design is the lack of a good criterion for deciding when we have a good measure. Where it is not feasible to validate the measure against an objective criterion, the following procedure is recommended. The instrument should be tried on a number of cases and the materials obtained should be coded by a number of judges. The judges are instructed about the meaning of the variables and it is their task to make independent assignments of the interviews or behavior records to ordered categories expressing amounts of the variable. Or a number of observers can use a behavioral scale in an actual situation and then check on whether they can get the data needed and whether they agree on the way in which they recorded the data. When this type of procedure is used, it soon becomes evident whether or not the question or device is bringing out the kind of material which can be reliably coded as satisfying the variable. This, of course, is not a true validity check, but it creates a presumption of validity and moves a long way toward developing the sort of measures necessary for the study. And the major codes for categorizing the interview and observational materials should be worked out during this pretesting stage. The final coding process should produce only minor additions to and revisions of these major categories.

In most surveys and field studies, the time and attention given to this form of pretesting are inadequate. And yet, to the extent that we are seeking to discover relationships there is no point of greater critical significance than the translation of the research objectives into operational measures. The most brilliant theory will yield no results and the most refined statistical analysis will be a waste of time if we fail to develop the operational measures of the variables with which we are concerned. Unfortunately, it is too often true that less time and effort are spent upon this problem than upon almost any other phase of the research process.

It is important to carry out the pretests of measures and procedures upon a population as similar as possible to the people who will be studied. If the field study is directed at a large group or community, it is possible to pretest the instruments without unduly influencing the results of the larger study to come. It is advisable,

in a situation in which the population is very large, to determine the sample in advance so that the pretesting can be carried on outside of the sample to be used.

In the pretesting of instruments and procedures, it is not essential to obtain a representative sample of subjects but it is important to try to include some of the main types of people who will be included in the final study. A pretest of an instrument on a few subjects from only the upper education brackets will not anticipate the problems to be met among those with minimum schooling. This, of course, applies more to the last stages of pretesting. Often, when we are first trying to develop a measure, we are less concerned with replicating the final field situation than we are with getting insight into the fundamental character of a given variable.

The pretesting of measures on populations very similar to those which will be used in the larger study is necessary to determine both the specific form of questions and of observational codes and the types of measures applicable to specific groups of people. It has been demonstrated that the written questionnaire is much more widely applicable than was formerly assumed. It has the great advantage over the personal interview of anonymity, which in some situations can compensate for the knowledge gained from actual personal contact with the respondent. Moreover, there are great economies in the use of written questionnaires as against interviews. But it is still true that written forms are much better adapted to well-educated groups. The lower skill levels among workers not only may have little formal schooling but they may have spent very little time with paper and pencil materials since their school days. It is more natural for them to express their ideas orally than in writing. The exact line to be drawn in the use of interviews as against written questionnaires can be determined through pretests. It is often advantageous, when employing written forms on a large scale, to draw a subsample of respondents for personal interviewing. Thus the biases in the one method can be checked by means of the other.

The Full-scale Field Operation

Ideally, almost all the research problems are solved before the study goes into the field, but unfortunately these problems do not

remain solved during actual field operations. The content of a research instrument, well pretested, takes on a new meaning because of its relation to events which have suddenly changed. The scales developed in advance for behavioral observation prove useless for bringing out the critical points in the group meetings under study. It is well, therefore, to have a research team in the field which has *some balance* between ingenuity and soundness of judgment. It is often necessary to improvise and make changes in procedures during the course of the field work. But if ingenuity in meeting new situations runs unchecked, the research team will lose sight of the original research objectives. In the actual field situation the pressure is strong to meet the practical needs of the moment at the sacrifice of long range research plans.

The skills and personnel for the field operation differ considerably from the requirements of a large scale survey. The tasks for the field worker are more varied and often more difficult than for the production interviewers in the survey. Not only must the field worker be able to enlist the cooperation of all groups in the community but much more of his interviewing must be with top leaders and ranking officials (10). In the Human Relations Program of the Survey Research Center, in which industrial and governmental organizations were studied, it was found that the interviewing staff for national surveys was not equipped to deal with the leadership levels of importance to the investigation. Field workers needed some experience with and knowledge of organizational and administrative problems. Individuals who had had considerable graduate training in psychology and the social sciences and considerable experience in industry and government proved to be the most effective workers. Moreover, field studies of this sort need to differentiate in the skills of their field workers and to include a few people who are capable of meeting outstanding leaders at something approaching their own level.

Since most field studies extend in time, controls should be set up to ensure comparability of the information obtained during different periods of the study. People interviewed at a later period in the study may have been affected by natural social events occurring since the start of the study or by reports from and interaction with earlier subjects. These two types of effects call for different procedures. The effects from the study itself suggest as rapid a

measurement of an entire group as possible and then passing on to another group. This is also the more economical method. But the effects of events in the world suggest that we do not interview all of one type of respondent at a single time period. The difference between groups may then be a difference between the people before and after an event. This conflict can be solved by concentrating heavily upon one group at a time but providing for small sub-samples of all groups for the major time periods of the study, even if it means some reobservation of the first people studied.

The field study must face the important problems of obtaining cooperation from the many individuals and subgroups in the structure under investigation. In a survey, it is generally not necessary to give much time or effort to obtaining the cooperation of respondents, since most individuals will give an hour or more to the interviewer, if he will come at a time convenient to them. But in a field study where the research workers may spend weeks and even months in the same community, where they may reinterview the same people, seek access to privileged information, and attend all types of meetings, the matter of obtaining community or group support takes on unusual importance (10). The following procedures are recommended as deserving consideration.

- 1 There is a real economy in going to the very top of the structures under study to obtain the cooperation of the ranking leaders. This is especially important if we are dealing with a hierarchical structure, as in industry, where lower levels are always dependent upon their superiors and are too insecure to risk welcoming investigators from the outside. It is very probable that the matter of cooperation will be referred up the line sooner or later, and there is a much better chance for a favorable decision if top leaders are consulted at the outset. In addition, the top leaders are often more likely to understand what the research is all about and are more open to conviction. The greatest resistance is frequently found among the petty officials, not only because of their insecurity but because of their general limitations.

- 2 The easiest entrance into a community or a social structure—namely, coming in as the ally of individuals or groups who have a special interest to exploit and who see the research as a means to their ends—should be avoided. Thus, the partisans of a specific reform in a community or the executive vice president of an indus-

trial company seeking information on the delinquencies of his subordinates may welcome the researcher and offer him support. The alliance with such special pleaders is neither ethical nor wise. The researcher's aim should be to enter the situation in the common interests of all parties, and his findings should be equally available to all groups and individuals. The cooperation offered a partisan has two disadvantages. It may result in a lack of cooperation on the part of other people, and it may exert undue influence upon the research objectives.

It may be of course, that this ideal is too difficult to achieve in practice, since the researcher may sometimes have to accept help where he can find it. But the broader the basis of his support, the more all groups in the community see that social science has potentialities for helping all those who want to avail themselves of its findings, the sounder our operations can be. To ensure this type of general community support, it is often advisable to set up an advisory, sponsoring council representing all the diverse interests in the community.

3 As much information about the study and its basic purposes should be given to the people of the community, as well as to the leaders, as can be revealed without prejudicing the results of the investigation. This problem of the amount of information which it is desirable to give is one of the real obstacles in field research and field experimentation. If the study is wide in its coverage, and if it extends over more than a few days, it is impossible to maintain a conspiracy of silence. An attempt to preserve secrecy merely increases the spread and wildness of the rumors. Yet the researcher does not want his potential subjects to know too much about his specific hypotheses and objectives. A common solution is to present an explicit statement at a fairly general level with one or two examples of items which are not crucial to the entire study.

In presenting an explanation of the study, it is well to utilize the accepted channels for communication in the various groups in the community. If the information is limited to a single channel, the study may become identified with the interests associated with that channel. In industrial studies an account of the coming research should appear both in the company house organ and in the union newspaper.

It should be recognized that any field study entails some risks. The introduction of observers prying into various aspects of community life may not be relished by all groups in the community. Certain questions may be offensive to some people. And at times the climate of opinion may be such that people will cooperate neither in launching the study nor in allowing themselves to be interviewed. The risks should be known from the scouting explorations and calculated in advance. In general, however, these risks are exaggerated in democratic societies. People in general like to be interviewed and are receptive to the idea of cooperating with social scientists. The general experience is that the usual practical obstacles in getting information which obsess the armchair critic are not so difficult to overcome as the problems of research design.

4 Another aspect of long run cooperation is the ethical standard of the research worker in keeping faith with people who have helped him. This involves a religious preservation of the anonymity of the respondent and a thorough carrying out of the spirit and letter of the obligations incurred in the study. Every provision should be made to protect the identity of the respondent. Analysis of materials should not be carried to such a point and case materials should not be quoted in such a fashion as to permit identification of specific individuals. If people are told that the identifying marks on their questionnaires will be removed, these marks should be removed. Research workers have rightfully made a fetish of this preservation of the anonymity of respondents and the assumption is that the rule holds save where there is explicit prior understanding and willingness on the part of the subject to become known.

Another, more subtle problem concerns the specificity with which findings are reported for particular groups and subgroups. Even if individuals are not identified, the reporting of results for subgroups can place them in positions of advantage or disadvantage. Here the general rule is that the cross breaks should be reported so as to show the general relations rather than to show the incidence of types of values and behaviors in small subgroups. Such general findings should be available to everyone. The researcher cannot, of course, control the uses to which his results are put, but he should attempt to provide equal access to his general findings for all groups.

The Analysis of Materials

In a well planned study in which hypothesis testing is the principal objective, the major codes will have been developed in the pretesting of instruments. In an exploratory study, many of the codes must still be developed after or during the period of data collection. Even though the study is exploratory, it is important to develop codes with some degree of conceptualization rather than to code for every phenotypical type of response (see Chapter 10). The major error of the novice is his fear that he will lose some of the richness of his materials if he does not cover all phenotypical details. In general, the more elaborate and the more detailed the coding the less useful it is in analysis. For analysis purposes, the data have to be recombined and the minute observations neglected.

A first step in analysis is to obtain frequency distributions on all measures. These straight run tables have several purposes. In obtaining these frequency tables, we can check for errors in the mechanical process of transferring material to Hollerith cards (all totals and subtotals can be checked, a process known as card cleaning). A second purpose is to give more information about the cross breaks which should be run. If a given frequency table shows no real spread of responses there is no point in attempting to correlate it with another variable. If, for example, our coded observations of the amount of hostility directed against members of the group show that 90 percent of the individuals had a zero hostility score, our chances for obtaining relationships with other variables are small indeed. If the total number of cases is very large, it is possible to work with such undifferentiated distributions. The analyst must always consider the size of the cells he will have in his correlational table before going ahead with cross breaks. Through an inspection of straight run tables decisions can be made about the combination of steps within a category or even about the combination of categories to comprise an index. The coding, for example, may have been too refined for the number of cases and the straight run table indicates that the seven degrees of approval may have to be reduced to five or to three. Or a number of different measures of participation in group activities may have to be combined into a single index. Finally, the frequency distribution needs to be examined in relation

to the correlation measure to be employed in that correlation assumes definite properties with respect to the distribution of the measures being correlated

The decision about the cross breaks to be employed is already determined in the hypothesis testing study. In the exploratory study, it is wise not to try to run every variable against every other variable but to decide on the basis of all the information available what the most promising relationships are likely to be.

In general, material from the observations and interview responses in a field study or a survey lend themselves more readily to correlation analysis than to analysis of variance. Analysis of variance is better suited to experimental studies where we have approximately equal groups assigned to known values of the experimental variable. The procedure of the breakdown or cross break is a correlational procedure, and every table reporting a cross break is a correlational table. Frequently, correlations are not computed, since the comparison of groupings within the table can show the relationship though the precise amount of relationship requires a correlational computation. Although it is a great advantage to be able to state the degree of relationship for the entire table by a single correlation coefficient, there is often no statistical rationale for such a procedure. Correlations make assumptions which the data may not justify. To compare the significance of the difference between two subgroups in the table demands fewer assumptions about the nature of the data.

The exploratory study calls for ex post facto analysis in which we want to make the most plausible interpretation of our findings even though the interpretation can never be definitive. The usual procedure here is to try to develop a relationship or to check on its spuriousness by holding factors constant through the use of triple breaks. The most ingenious and systematic use of this method is exemplified in the work of P. Lazarsfeld who with P. Kendall has presented a codification of such analysis procedures in *Continuities in Social Research* (13). Three main types of elaboration are described which differ not so much in essential method as in giving the analyst a framework for thinking about the relationships in his data. One type of elaboration is to control for spurious factors. In this case, we have discovered a relationship between two variables but we want to make sure within the limits of our data that it is

not a function of a third factor which would explain away the relationship. Thus, a relationship which we discover between job satisfaction and skill level of the job may be due to the fact that higher skill jobs are better paid. A triple break will allow us to hold wages constant for different skill levels in relation to job satisfactions.

Instead of trying to explain away our findings as due to some spurious factor, we may want to show that some other factor, whose timing we can establish, has intervened to produce the relationship. Thus in the studies of the Research Branch of the Army, the poorer educated men, though in general less critical, were more likely to feel that they should have been deferred than were the better educated men (13). Lazarsfeld hypothesizes that a variable may have intervened between the factor of education and feelings about deferment as the real reason for the relationship—specifically, the factor of relative deprivation. The poorer educated men were supposedly coming from groups where deferments were higher than among the better educated. If data were available on the rate of deferment in this study and if, on the triple break, the correlation was really with deferment rate, then this interpretation of the relationship becomes plausible. The essential point in this type of elaboration is that the test factor can be pinned down in time to some actual variable that has intervened.

A third type of elaboration is specification—that is, trying to find the conditions under which a relationship will be accentuated. We may find in an industrial study, for example, a positive but small relationship between good human relations skills of the foreman and the morale of his workers. We have theoretical reasons for expecting the relationship to be higher, and hence we seek to specify the conditions under which the correlation should increase. We postulate, therefore, that in departments where the foremen are effective in dealing with their superiors, the morale of the men will be higher, since, in addition to understanding their men, they can accomplish things on their behalf. By means of a triple break in which we compare the morale of workers under foremen with good human relations practices with the morale of workers under foremen less skilled, for different degrees of conditions of foreman effectiveness up the line we can document our specifications about the nature of the relationship.

These three types of elaboration are convenient ways in which the researcher can utilize post hoc analysis in making his data yield

the most meaning. In this process, the findings are used to suggest hypotheses, which are then examined and tested further through manipulation of the data. The form of statistical manipulation in the three types is the same, but the logical and theoretical implications differ. It is true, of course, that this type of hypothesis testing after the fact is limited by the data already gathered and is therefore no substitute for the controlled experiment. Nevertheless, such analysis can greatly increase the probability of the soundness of the interpretation of the data.

The meaning of the data should not be speculated about when it is possible to test the speculations in the data themselves. Often by cross breaks and by holding factors constant the statements which the researcher sets forth as interpretation can be checked against his own data. If confirmed, it can be stated not as speculation but as a finding within the limitations of his study.

Field studies frequently pose the problem of the proper N to use for computing the statistical significance of a difference. It is common to use the number of individuals in a subgrouping as the N without regard to the possibility of clustering effects. Where we are dealing with a cluster, the proper N to use in computing the statistical significance is the number of clusters, not the number of individuals. Suppose, for example, we are studying the effects of different types of adult education practices upon the acceptance and use of local public health clinics. There are three discussion groups of 12 members each in which a participation method was used and three groups of 50 members each in which a lecture method was used. In comparing the effectiveness of these methods, it is not justifiable to regard our N 's as 36 and 150, respectively. We do not have 36 separate and independent manifestations of the discussion method since in any one group the use of the method was colored by the discussion leader and the composition of the group. Hence the true N here is 3. Similarly, in the lecture method there are not 150 independent effects of the method but rather three major applications for it. It should be remembered that the probable error of the difference is based upon N 's which are independent measures of the effect in question.

The field study is particularly susceptible to cluster effects. In the laboratory we can set up our experiments to guard against this difficulty. In surveys we can sample in random fashion to avoid clustering. But in a community or a group under field study our

cases are often pocketed in homogeneous subgroups, where the clustering effect must be carefully considered

THE PLACE OF FIELD STUDIES IN PROGRAMMATIC RESEARCH

In building a science of social psychology it is desirable to take advantage of the different settings in which our phenomena occur and of approaches which utilize the particular advantages of a given setting. The field study is unique in enabling us to observe and measure social processes in their natural occurrence. On the one hand it can give depth of understanding to survey findings. On the other hand it can give to the experimenter rich insights and hypotheses for more rigorous experimentation and can prevent the laboratory from developing a system of concepts which have little to do with the way in which people really behave.

In dealing with social events in a natural setting the field study must operate in an open system of interacting variables in which so called alien factors may influence the outcome. But the limits of our subject matter are as yet so poorly defined that not all psychologists would agree about what should be considered an alien factor. Hence at the present stage of our discipline, there is much to be gained from working both in the laboratory, where the restricted model from physics can be the ideal, and in the field situation, where the looser model sometimes employed in the biological sciences is the prototype.

Another way of describing the potential usefulness of field studies is to call attention to the history of the psychology of perception. Early experimental work, in its attempt at scientific rigor, ignored the wholistic qualities of the psychological field and proceeded in overanalytic fashion to work with highly artificial elements of consciousness. The same weakness was to be observed in the early experimental work on social processes in the laboratory in the emphasis upon social facilitation. It is important to analyze the complexities of social interaction and social relationships, but there is always the danger that our analysis may miss the significant functional properties and deal with elements which are more artifacts than functional entities. Field studies, by their close con-

tact with on-going social events, can serve as a check against the omission of significant variables. In a long-range program of research, there should be a two-way interaction between experimentation and field studies. The findings from field studies may raise questions which need for their solution the more rigorous methods of experimentation. Conversely, the experimental results can furnish a basis for the formulation of some of the research objectives of the field study. If the relations which hold within the laboratory also prove of significance in the field situation, it is an indication that the laboratory has not moved in too artifactual a direction.

There is an important intervening step, however, between the field study and the laboratory experiment—namely, the field experiment. The natural experiment, previously discussed, is a social change which takes place without any action by the social researcher. It just happens to be an interesting change for him to measure. The field experiment, however, is a social change engineered by the researcher. He is instrumental in effecting the manipulation of a set of variables in a life situation. The field experiment is the logical connection between the field study and the laboratory experiment. It follows the logical procedures of the laboratory more closely, but it nevertheless deals with factors operating in the field situation and thus, like the field study, is more concerned with "global" variables.

Field studies can furnish the essential information which will make a successful field experiment possible. The most natural step in following up the leads from the field study is experimentation in the same situation, so that there can be some actual manipulation of the factors which appeared to be causal determinants in the field study. It is to field experimentation, therefore, that the next chapter is devoted.

BIBLIOGRAPHY

1. Adorno, T. W., *et al.* *The authoritarian personality*. New York: Harper, 1950.
2. Barnard, C. I. Functions and pathology of status systems in formal

- organizations In Whyte W F *Industry and society* New York McGraw Hill 1946 pp 46-83
- 3 Child I L *Italian or American?* New Haven Yale Univ Press 1943
 - 4 Curle A and Trist E I Transitional communities and social reconstruction Part II *Hum Relat* 1947 240-288
 - 5 Davis A Gardner B B and Gardner M R *Deep South a social anthropological study of caste and class* Chicago Univ of Chicago Press 1941
 - 6 Dollard J *Caste and class in a Southern town* New Haven Yale Univ Press 1937
 - 7 Festinger L Schachter S and Back K *Social pressures in informal groups* New York Harper 1950
 - 8 Harbison F H and Dubin R *Patterns of union management relations* Chicago Science Research Associates 1947
 - 9 Hollingshead A B *Elmtown's youth* New York Wiley 1949
 - 10 Jacobson E Kahn R L Mann F C and Morse N C (eds) Human relations research in large organizations *J Soc Issues* 1951 7 No 3 175
 - 11 Jones A W *Life liberty and property* New York Lippincott 1941
 - 12 Katz D Survey techniques in the evaluation of morale In Miller J G *Experiments in social process* New York McGraw Hill 1950 Chap 5
 - 13 Kendall P L and Lazarsfeld P F Problems in survey analysis In Merton R K and Lazarsfeld P F (eds) *Continuities in social research* Glencoe Free Press 1950 pp 133-196
 - 14 Leighton A *The governing of men* Princeton Princeton Univ Press 1945
 - 15 Lynd R S and Lynd H M *Middletown* New York Harcourt 1929
 - 16 Malinowski B *Crime and custom in savage society* New York Harcourt 1926
 - 17 Newcomb T M *Personality and social change* New York Dryden Press 1943
 - 18 Roethlisberger F J and Dickson W J *Management and the worker* Cambridge Harvard Univ Press 1939
 - 19 Schank R L A study of a community and its groups and institutions conceived of as behaviors of individuals *Psychol Monogr* 1932 43, No 2
 - 20 Stouffer S A Suchman E A DeVinney L C Star S and

- Williams, R. *The American soldier*, Vol I Princeton Princeton Univ Press, 1949
- 21 United States Strategic Bombing Survey, Morale Division *The effects of strategic bombing on German morale* Washington U S Government Printing Office, 1947
- 22 Warner, W. L., and Lunt, P. S. *The social life of a modern community* New Haven Yale Univ Press, 1941
- 23 Warner, W. L., Meeker, M., and Eels, K. *Social class in America* Chicago Science Research Associates 1949

Experiments in Field Settings

John R P French, Jr.

The desirability of widening the scope of experimental social psychology through experimentation in "real life" settings has been recognized for some time. Lewin has pointed out that "Although it appears to be possible to study certain problems of society in experimentally created, smaller laboratory groups, we shall have also to develop research techniques that will permit us to do real experiments within existing 'natural' social groups. In my opinion, the practical and theoretical importance of these types of experiments is of the first magnitude" (30, p. 164).

WHAT IS A FIELD EXPERIMENT?

Although considerable progress has been made in developing such methods, the field experiment is not yet a well developed method of basic research in social science. Rather, there is a variety of related methods, such as action research, evaluation research, operational research, etc., which may include experimental studies in field settings. In discussing current conceptions of experimental method in sociology, Greenwood (18, Chap. 2) describes five types: (1) the pure experiment, which is here called the laboratory experiment (see Chap. 4), (2) the uncontrolled experiment, or what we have called the natural experiment (see pp. 78-79), (3) the ex

post facto experiment, in which the investigator tries to trace back wards from an effect which has already occurred to its causes, (4) the trial and-error experiment, which seems to refer to all sorts of trials by laymen of new forms of social behavior, (5) the controlled observational study. None of these five categories describes the field experiment, although the major dimensions along which the categories differ are fairly adequate for such a description. Accordingly, the meaning of the field experiment and its relation to other methods can best be clarified by considering variations of three aspects: the design of the research, the setting, and the purpose.

The essential factor which distinguishes the field experiment from the more common "field study" is the *design of the research*. The field experiment involves the actual manipulation of conditions by the experimenter in order to determine causal relations, whereas in the field study the researcher uses the selection of subjects and the measurement of existing conditions in the field setting as a method of determining correlations. As we shall see later, one of the crucial methodological problems of the field experiment is devising ways of manipulating the independent variable. It is this difficulty which has led to the use of the "natural experiment," in which the researcher opportunistically capitalizes upon some on going changes and studies their effects in an experimental design. If these natural changes have already occurred by the time the social scientist arrives on the scene, it may still be possible to gather sufficient data after the fact to fill out the design of a crude ex post facto experiment. In the field experiment, the manipulation of the independent variable is not left to nature but is contrived, at least in part, by the experimenter, thus, the design is planned by him beforehand.

The wide variety of truly experimental studies that have taken place in field settings vary in *purpose* all the way from the development of social psychological theory to the immediate solution of some practical social problem. Sometimes both purposes are present, as in the experimental forms of "action research" which have the dual purpose of bringing about social change and at the same time contributing to basic social science (7, 8, 27, 39). But most applied research has the major purpose of obtaining facts and attitudes or evaluating methods which will be of immediate value in solving some specific applied problem, although theoretical development may be a minor purpose (44). Such practically oriented research is

the most common type of field experiment. For example, studies evaluating the relative effectiveness of two types of political propaganda or of two teaching methods or of several advertising appeals attempt to get fairly immediate answers to practical problems of politics, education, or advertising without attempting to apply any general theory. In fact, there are very few field experiments with a strongly theoretical orientation, yet there is a sufficient stockpile of theory and knowledge in social psychology to warrant many such field experiments. Most of all we need experiments whose purpose is to test the applicability to real life situations of the known scientific laws or hypotheses specifically developed in controlled laboratory settings. Throughout this chapter we shall try to emphasize the scientific and theoretical purposes of the field experiment.

The *setting* for a field experiment is some real existing social situation in which the phenomena to be studied are commonly found. By implication it is not an "artificial" situation created in a research laboratory. This distinction, however, is not nearly so clear cut as it seems at first glance, for it is not important whether the social phenomena occur in a building called a laboratory rather than in a school or some other real social institution. The relevant distinction here seems to be between studying real and studying artificial social phenomena. One meaning of 'artificial' as applied to the behavior of people in the laboratory seems to be that their behavior is determined by their role of being a subject, that they would not act the same way if they were not in this role. Block and Block have pointed out that the middle class subject almost invariably structures the experimental situation as one calling for a submissive role in relation to an authority figure (3). In so far as social behavior is role determined it is clear that findings obtained with one role cannot be generalized to apply to other roles without additional research. In addition the behavior of subjects in a laboratory experiment is highly restricted by the rules and procedures instituted in order to control conditions. Frequently, this simplification involves the creation of new groups which will not be influenced by their past history or their present social setting. The laws which hold for such restricted situations may not apply without changes to the more complex settings of real life. Usually a field experiment is not subject to such artificiality and thus avoids this problem of generalizing to real life situations. That this is not always the

case, however, is well illustrated in the famous Hawthorne experiments (37). From a methodological point of view, the most interesting finding was what we might call the "Hawthorne effect." In order to manipulate more precisely the physical factors affecting production, the experimenters had set up a special experimental room for a small group of girls who were wiring relays. This wiring room was separated from the rest of the factory, and the girls working in it received special attention from both outside experimenters and the management of the plant. Careful studies of this wiring group showed marked increases in production which were related only to the special social position and social treatment they received. Thus, it was the "artificial" social aspects of the experimental conditions set up for measurement which produced the increases in group productivity. The distinctions between "artificial" laboratory experiments and "real" field experiment are, therefore, matters of degree; one field experiment will vary greatly from another in the complexity of the social setting, the controls introduced by the experimenter, and the roles played by the subjects.

For the purposes of this chapter, then, we shall define a field experiment as a theoretically oriented research project in which the experimenter manipulates an independent variable in some real social setting in order to test some hypothesis.

PLANNING A FIELD EXPERIMENT

The idealized sequence of steps in planning a scientific experiment might include, first, the selection of a problem on the basis of theoretical considerations and the formulations of precise hypotheses, and then the selection of some appropriate methodology and the creation of an experimental design. No such simple sequence is usual in planning a field experiment, because the social scientist does not have free access to the whole world as his laboratory or the power to carry out most of the experimental manipulations he might conceive. Accordingly, he must proceed somewhat opportunistically. Frequently, he is first presented with an opportunity to do research in some specific setting, so that the choice of problem and method will come afterwards. In most cases it will be wise to consider simultaneously the selection of a problem, the selection

of a field setting, and the designing of the study, since all three of these factors are highly interdependent. The kind of control groups that are needed, for example, will depend on the problem selected, and the possibility of obtaining these groups will depend on what is available in the setting. This interdependence should be kept in mind as we discuss, in sequence, selecting and formulating a research problem, selecting a setting, and research design.

Selecting and Formulating a Research Problem

SELECTING THE PROBLEM The research problem and the research methodology should be selected for their appropriateness to each other. It is not easy at present to make generalizations about the kinds of problems for which the field experiment is an appropriate methodology because there has been too little field experimentation in social psychology. Yet one can see certain common types of problems in the work that has been done. Perhaps the largest category of problems for which a field experiment has been the appropriate methodology would be the studies on "how to do it." The field experiment has been used for studying a variety of techniques and methods, such as advertising techniques, training methods (4, 19, 31, 38), the effects of group decision (28, 29) and participation (9), methods of group therapy (10), political propaganda (1, 20) etc. In all of these cases, one or more methods for bringing about a change already existed prior to the research. In other words, these experiments tested the effectiveness of change procedures which had been used for purposes other than research.

Also, there have been several experiments which made use of the existing formal or informal social structure. Jackson, for example, studied leadership by actually exchanging foremen (21). Snyder made use of existing leaders and their groups in experiments on the influence of military leaders (40). Van Zelst studied the effects of sociometric regrouping on the productivity of work groups (43). Social structure has also been used in experiments on rumor (2).

Field experimentation seems also to be an appropriate method to use with a problem or phenomenon too difficult to study in the laboratory—for example, the changing of food habits (28) or the effectiveness of political propaganda (1, 20) or "mass education" (42). Probably a field experiment is used in many of these cases not so

much because it would be impossible to devise an experiment in the laboratory but because such artificially created phenomena would be too weak or would not be dynamically equivalent to the real thing. Thus, Annis (1) collaborated with the printer to have an editorial "planted" in the university daily so that the subjects would read it with their normal set without knowing that it was an experimental stimulus. Had they known, it is doubtful that the experiment would have been dynamically equivalent to the usual political propaganda.

Conversely, one of the more important determinants in selecting a problem for a field experiment is the fact that it is possible to study by this method only those hypotheses in which one can manipulate the independent variable or produce a change. Here it is most important to keep in mind the rule: "*Start strong.*" Except where experimental methods are already well developed, the experimenter should attempt to maximize the variations in the independent variable or the differences among the experimental treatments. This is necessary both because our methods of measurement in social psychology are often so crude as to be able to reflect only rather gross changes and because the experimenter, at least at the beginning, is quite likely to produce far smaller changes than he attempts. There are a variety of determinants of the changes that can be produced, such as the role of the experimenter, his knowledge, and his skill. These will be discussed further in a later section of this chapter.

This same rule would require that we study major rather than minor independent variables. If our dependent variable is determined by several factors, we cannot hope to determine experimentally the effects of the minor ones until we understand (and can control) the major ones which account for most of the variance.

In other cases, the field experiment seems to have been the preferred methodology because it reduces the problem of generalization and application of results. Much of the "operational research" in the armed services and in other organizations is of this character. Such applied research usually evaluates the effectiveness of operational procedures. Many a study of college sophomores has been criticized on the ground that the same results would not hold for soldiers or industrial workers, but such an objection cannot be made for a field experiment using soldiers or industrial workers as

of a field setting, and the designing of the study, since all three of these factors are highly interdependent. The kind of control groups that are needed, for example, will depend on the problem selected; and the possibility of obtaining these groups will depend on what is available in the setting. This interdependence should be kept in mind as we discuss, in sequence, selecting and formulating a research problem, selecting a setting, and research design.

Selecting and Formulating a Research Problem

SELECTING THE PROBLEM. The research problem and the research methodology should be selected for their appropriateness to each other. It is not easy at present to make generalizations about the kinds of problems for which the field experiment is an appropriate methodology, because there has been too little field experimentation in social psychology. Yet one can see certain common types of problems in the work that has been done. Perhaps the largest category of problems for which a field experiment has been the appropriate methodology would be the studies on "how to do it." The field experiment has been used for studying a variety of techniques and methods, such as advertising techniques, training methods (4, 19, 31, 38), the effects of group decision (28, 29) and participation (9), methods of group therapy (10), political propaganda (1, 20), etc. In all of these cases, one or more methods for bringing about a change already existed prior to the research. In other words, these experiments tested the effectiveness of change procedures which had been used for purposes other than research.

Also, there have been several experiments which made use of the existing formal or informal social structure. Jackson, for example, studied leadership by actually exchanging foremen (21). Snyder made use of existing leaders and their groups in experiments on the influence of military leaders (40). Van Zelst studied the effects of sociometric regrouping on the productivity of work groups (43). Social structure has also been used in experiments on rumor (2).

Field experimentation seems also to be an appropriate method to use with a problem or phenomenon too difficult to study in the laboratory—for example, the changing of food habits (28) or the effectiveness of political propaganda (1, 20) or "mass education" (42). Probably a field experiment is used in many of these cases not so

accounting for the distinctive characteristics of role playing—namely, the degree of involvement in the role. Other characteristics were neglected, and attention was focused on this one variable. Accordingly, the experiment attempted to create three degrees of involvement in the role in order to test explicit hypotheses concerning the effect of such involvement on perception, feeling, and participation in discussion.

As a problem or a social technique for bringing about a change becomes more conceptualized, it also becomes clearer that it involves several hypotheses and that there is a real choice as to how global or how analytical to be in stating a hypothesis. Whether to work with a global syndrome or with a narrowly defined variable poses a problem about which there is considerable difference of opinion among different disciplines of the social sciences and among different theoretical approaches. We cannot discuss it fully here. Rather, we shall attempt to point out certain aspects especially relevant to the field experiment.

Consider Fig. 1 as a schematic representation of the causal interrelations among the experimental manipulations, the independent variables to be manipulated, and the dependent variables whose changes are to be measured. In this example, our general hypothesis predicts that the more democratic the behavior of the first-line foreman, the higher the production of the employees (the arrows indicate the causal relations). Our independent variable of foreman behavior might be manipulated, for example, by training the foremen (A in Fig. 1). But how shall we define "democratic behavior" for the purposes of our training course? This pattern of behavior consists of many parts or dimensions which might be treated as separate variables; for example, the amount of freedom or closeness of supervision (I_a), the amount of participation permitted in planning (I_b), the arbitrariness of discipline (I_c), and how much the foreman does to satisfy the needs of his employees (I_d).

But these aspects are interdependent so that a change in one will produce a change in any neighboring aspect. (In Fig. 1, low interdependence is represented by a heavy boundary and high interdependence is represented by a less heavy boundary between neighboring regions.) Now, in analyzing the data from a nonexperimental field study of production as determined by the behavior of the foreman, the investigator is free to choose any part or any combina-

subjects. In short, by study of the *specific* setting and problem which is to be solved it is possible to avoid the problem of generalizing from one situation to another.

On the other hand some problems are appropriate for a field experiment precisely because they require generalization to a real situation. These are the hypotheses proved in laboratory experiments but not yet sufficiently studied in field settings. We have already indicated that this should be an increasing source of field experiments.

Finally, of course, many problems appropriate for field experiments have come out of field studies. Often these are theoretical problems in which a hypothesis has been too confounded in the nonexperimental studies or where a correlation has been established without the possibility of determining the direction of causation. The Survey Research Center at the University of Michigan has conducted several field studies of industrial organizations in which it has found that organizational units with high productivity have more democratic supervision. Because either could be reasonably interpreted as the cause of the other, the Center put the matter to experimental test (34-35). In a field experiment the locus of organizational control was actually manipulated toward more decentralization in one set of groups and toward more centralization in another matched set of groups. The hypothesis would predict increases in productivity in the first treatment and decreases in the second treatment. With this experimental design it is no longer possible to say that high productivity might be causing the democratic supervision.

FORMULATING THE PROBLEM In order to make a scientific contribution the hypothesis selected for study in a field experiment must be statable in general terms—i.e. the variables must be conceptualized (30 Chap. 2). This will tend to be true almost automatically for an experiment designed specifically to test some law which was developed in basic research. If the initial selection of a problem has been more practically oriented, however, it will be important to try to conceptualize the problem or the procedures. For example, Rosenberg's first step in an experiment on the effectiveness of role playing as a method of training (38) was to try to develop some insight into the nature of role playing. This led to the identification of one major variable which seemed important in

tion of parts of foreman behavior as the independent variable this includes the freedom erroneously to treat I_a plus I_d as a single variable instead of choosing some combination more closely corresponding to the actual interdependence (for example, I_c and I_d) or choosing I_a alone. In a field experiment, on the other hand, there is less probability of error in deciding how analytical to be in formulating the independent variable.

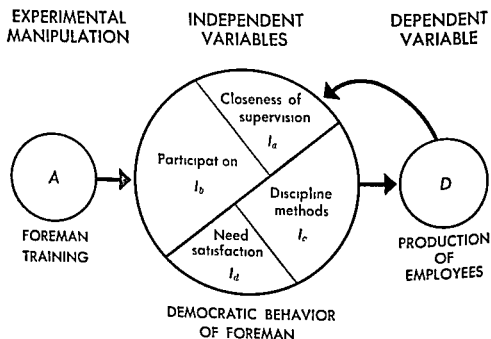


FIG 1 The causal relations among experimental manipulations independent variables and dependent variables

In planning the foreman training let us suppose that the experimenter tried to be too atomistic by varying I_a alone. If, in fact, I_a is highly interdependent with I_b , then a change in I_a would produce a change in I_b . Thus, the experimenter would have actually varied both variables—a result which corresponds better to the real structure of the situation. If he has measures of all four independent variables, the experimenter would then notice the changes in I_b and would reformulate his independent variable in a way which would

yield better predictions. By attempts actually to manipulate a variable in the field setting, it seems that the social scientist has a better chance of finding out what parts necessarily hang together. He partly avoids the error of a too broad or too narrow definition of his variables. This is important, because it is one way of making our concepts conform to the essential nature of the phenomena we study.

Fig. 1 illustrates also the advantage of the field experiment over the field study in determining the direction of causation. The field experiment can distinguish between independent and dependent variables, whereas the field study can only establish correlations among variables. In this example we have assumed that increasing the level of production will also change certain parts of the democratic behavior of the foreman—*i.e.*, it will reduce the closeness of his supervision. Since we have also assumed that closeness of supervision causes lower production, we have here a circular social process. In order to analyze fully the nature of this circular process, one would have to conduct two field experiments in which each of the two variables was manipulated as an independent variable. Probably a good many real social processes contain circular causal relations of this type. It is important to analyze these phenomena experimentally.

In formulating a problem for a field experiment, one should state explicitly the hypothesis to be tested. As in laboratory experiments, the more this hypothesis is derivable from a general theory, the better. It is also important to formulate hypotheses about all other possible causes of the same dependent variable so that one can try to hold these other factors constant in the design of the experiment.

RESEARCH OBJECTIVES AND PRACTICAL OBJECTIVES. In the selection of a problem for a field experiment, the range of possible problems will be determined partly by the relation of the theoretical purposes of the scientist to the practical purposes of the subjects and their institutions. Sometimes there will be commonality of goals; often there will be differences. These differences in goals mean that the activities in which the researcher must engage to reach his scientific objectives may be seen by the practitioner as interfering with his goals. Accordingly, these differing purposes set a limit on the freedom of the experimenter to choose any problem

of interest to him. The less interest the practitioner has in the development of science, the more limited will be the researcher's choice.

One tempting solution to this conflict, which may frequently have bad effects on the research, is the attempt to study too many problems at the same time—that is, the social scientist collects data in order to test certain theoretical hypotheses and at the same time tries to collect additional data to satisfy several of the practical needs of the organization in which he is working. This results in a dispersion of effort and a lowering of the quality of the research. Another solution which reduces the scientific value of the research is to compromise on the degree of generality in formulating the problem for the purposes of science, maximum generality and abstractness of the concepts are desirable, whereas the practitioner often wants knowledge about his specific situation with all its idiosyncrasies. The best solution to this problem seems to be to recognize frankly the possibilities of conflict as well as the possibilities of mutual interest and then to select for study a problem of maximum benefit to both the scientist and the practitioner. We shall have more to say about this cooperative relationship later (see pp. 123f).

Selecting a Setting

ESTABLISHING CONTACT Unless he is already established in the research department of some organization, the experimenter will face the problem of establishing contact with appropriate field settings. Social scientists in organized research centers will usually find this much easier because of the many institutional contacts which already exist. If the research center offers services to practitioners (consultation, training institutes and conferences, lectures, etc.) there will be many client organizations with which there is good rapport. There may even be frequent requests that the center do research projects.

In other cases the scientist must take the initiative to develop contacts. One lead to possible settings is an examination of the kinds of organizations already supporting similar research. Government and business organizations, for example, now support a variety of field projects. In other organizations not supporting research the

researcher will find some people who are favorably disposed toward social science research and willing to cooperate. Contacts can be made most easily with those organizations which feel the need for research to help solve their own problems. Particularly for a field experiment, which will entail certain new and unforeseen changes in the organization, it is best to contact organizations where there is *both* a motivation to contribute to science and a *realistic* expectation that the research will benefit the organization. This suggestion implies that the researcher must show his willingness to help the organization. No statements, however, will be as convincing as past services to organizations or past field studies which have had beneficial results.

In selecting a setting for a field experiment, the researcher will, of course, look first for a social situation which contains the phenomena he wishes to study. This is not a simple matter of determining the presence or absence of the phenomenon, but rather a question of discovering whether it occurs in a sufficiently pure form or strongly enough so that research is feasible. Actually this may require considerable knowledge and understanding of the setting. For example, a researcher who wished to study the strength of group standards about performance in work groups as a function of the cohesiveness of the groups selected bomber maintenance crews in the Air Force. This seemed an ideal setting because there were so many identical groups. Later he discovered that these groups were too small, that their performance was difficult to distinguish from the work of other maintenance men, that the crews were not in fact, identical, and that probably the group standards were generally weak.

THE INFLUENCE OF THE EXPERIMENTER Among the many settings which may be relevant to the problem he wishes to study, there will be important variations in the degree to which the experimenter has freedom and power in the setting or organization to be studied. First of all, he needs the power to manipulate some independent variable or to try out some technique of social change in a controlled design. The power and influence required for such experimenting will depend very much on how upsetting the experimental procedure is to the organization.

Both his power and his skill in carrying out an experimental treatment in a field setting may depend upon the researcher's rela-

relationship to the practitioner. If his own skills or the power of his role are insufficient, the researcher may be able to use a skillful practitioner as the actual manipulator of a variable. Not only does such a relationship increase the possible range of successful experimental problems, but there is frequently a great value in separating the action role from the role of data collector. For example, in experiments on therapy, training, and the like, there is probably much less bias if the researcher is not also at the same time the therapist or trainer. Thus it will often be desirable or even essential for the researcher to select a setting in which such help and cooperation from a skillful practitioner are available.

It is also desirable that the experimenter have the power to control other factors which might confound the experiment. In most field settings, this seems to be possible to only a limited degree. An inferior substitute for actual control is knowledge, especially foreknowledge, about the confounding factors. For example, if one wishes to conduct an experiment on production in a factory, it may be essential to prevent the factory from laying off workers in the middle of the experiment yet it may be impossible for the researcher to control this. However, if he can know far enough in advance of any impending lay offs, he may be able to plan his experiments during a time when this confounding factor will not disturb them.

Freedom of access to the data is an obvious consideration in selecting a setting for a field experiment. Organizations vary widely not only in their willingness to permit the scientist to collect data but also in the extent to which good quantitative data are already available. An equally essential freedom for the scientist is the freedom to publish the results of his research without undue censorship.

DIFFICULTIES IN THE FIELD SETTING Conducting field experiments in areas in which there is crystallized social conflict raises some extremely difficult and delicate problems. In order to develop an adequate and useful social science, it now seems essential to make field studies of such conflicts. This would imply the selection of settings in which tension and conflict are great. In order to study problems which are not a function of conflict, however, these settings should be avoided, because there is likely to be an iron curtain preventing study and the data which can be collected are most likely to be biased. In a setting in which there is strong hostility between management and labor, gaining the cooperation of one group may

virtually preclude gaining cooperation from the other (22). Thus, the scientist may be forced into a position which will be seen as "taking sides," and therefore access to important data will not be permitted. It seems important for the scientist to maintain an independent role even if it means that he does not study severe conflict situations (17, 24). It is especially true that too much conflict within an organization can make impractical the launching of a field experiment which requires a great deal of cooperative effort (22).

The extent of recognized problems in an organization and the felt need for help in solving these problems present a similar difficulty in selecting an appropriate setting. "If the organization is faced with a large number of acute problems it is not likely that its leaders will feel able or willing to divert efforts from 'fire fighting' to the investigation of the causes of their difficulties. On the other hand, the management which feels that it has solved all of its problems has little incentive to engage in a research relationship of any kind" (22, p. 66). Engaging in a field experiment requires more change than permitting a field study. Hence, an organization usually must perceive the problem to be more acute in the former than in the latter case in order to engage in research. Jaques (25), who places heavy emphasis on the therapeutic aspects of his work, studies only those problems concerning which the client requests help.

THE NEED FOR SCOUTING. Taken together, all the considerations so far discussed imply the need for a considerable body of information, some of it rather subtle and difficult to acquire, before a wise choice of setting can be made. Unless the experimenter already has direct knowledge of the situation through previous field research in it, it will be necessary for him to do a considerable amount of "scouting" (see also Chap. 2). The experimenter should visit the setting so that he may directly observe behavior and interview a number of people in different positions. It is particularly important to discuss their conception of the experiment with those who will help to make the decisions about the research and, if possible, to talk with those who might have conflicting attitudes toward it. Particularly where there is potential conflict between opposing parties, it may take longer to get agreement that the research should proceed than it takes to get the necessary informa-

tion about the setting. In his intensive therapeutic study of a factory, Jaques spent three months working out the arrangements for the study and the relations between the research team and various parts of the factory (24).

EXAMPLES A few examples will make clear the several aspects of field settings discussed above and will illustrate some of the roles and relationships that experimenters have established. We shall choose examples illustrating three main types of relationships—first, where a single experimenter is an outsider who belongs neither to the organization being studied nor to a research organization, second, where the researcher belongs to a large independent research organization, and third, where he is a member of a research unit within the organization studied.

Jackson's study of leadership (21) is an interesting example of the kind of cooperation a skillful experimenter can obtain even without the help of a large research organization. The experiment was performed as an M.A. thesis while the author was a graduate student. After discovering that a military setting was not appropriate for the design, he was introduced by his psychology professor to a vice president of a company which had sent representatives to the professor's training seminar. From then on the experimenter was on his own in developing rapport, obtaining approval of the project, and carrying it out. The experimental manipulation required exchanging the three best and the three worst foremen, an operation which involves more than the usual upset of the organization. In addition, it was necessary to prevent any further transfers of the six experimental foremen and their seventy men during the three months of the experiment. Accordingly, the experimenter stressed from the very beginning both the importance of these requirements and the fact that they might be difficult and upsetting at times. Then, with the help of the vice president, a series of meetings was held to obtain understanding and approval by the other managers involved. This type of clearance throughout management (and with the union) resulted in enough support for the project so the manipulation was carried out and conditions were held constant with respect to transfers, despite pressing reasons for making changes. The main reasons for success in obtaining such cooperation seemed to be the experimenter's skill in explaining the project, in conducting meetings and in leading group decisions, and the fact that the

company got as a by product, without cost, the methods it needed for evaluating foremen

An example in which the relationship is between an independent research organization and a client company is found in the work of the Survey Research Center at the University of Michigan. This research agency had a quite specific contract with a large insurance company to conduct a field experiment (see p 107) on the effects of the location of control and regulation (34, 35). This contract specified the independent role of the research organization and guaranteed its right to publish the findings. Since the research team had no authority in the company, it could carry out the experimental manipulation only through the help and cooperation of the management. Such a manipulation as delegating control and regulation, of course, required not only a long series of meetings to arrange the new duties and responsibilities of supervisors but also a long series of training sessions to prepare them to assume the new responsibilities. In this process the research team used both members of the company and outside specialists. To a great extent such a large, complex, cooperative venture could succeed because the company and the research organization had already cooperated on a field survey during which good relations had been established and the research organization had been able to discover in detail the initial state of the phenomena to be studied in the experiment—namely, the actual locus of control and regulation. In addition, however, they had acquired considerable information about the internal relations within the organization and the nature and acuteness of various problems.

Quite a different relationship between the experimenter and the industrial setting was involved in some field experiments at the Harwood Manufacturing Corporation (9). Here the author had an inside role as personnel manager and director of personnel research. Even the first studies in this organization could involve experiments, because the top management was favorable toward research and sophisticated about it. This was due to the unusual fact that the president of the company had a doctorate in psychology. The author's position in top management, and strong support from the president, permitted freedom to do experiments. There was some possibility of holding conditions constant during an experiment—for example, during experiments on the effects of group

decision on the level of production it was possible to prevent supervisors from exerting any influence on the production of their employees—and where it was not possible actually to control confounding factors it was usually possible to know about their existence in advance. On the other hand, this inside position meant that the experimenter was subjected to stronger influences of practical considerations of company goals and policies. Although this inside role led to better understanding of the dynamics of the factory as well as inside knowledge of events before they occurred, it also clearly limited access to certain information. Initially, when there was no union, it was rather easy to obtain a wide variety of information about human relations in the plant by interviewing people at all levels, later, shortly after the plant became unionized, it was impossible to establish rapport with some union members and consequently to secure valid information from them. For certain kinds of problems, therefore, it is important for the researcher to be an outsider with an independent role.

Research Design

The problems of research design for a field experiment are in principle the same as they are for a laboratory experiment, but the difficulties in manipulating and controlling variables are often greater in the field setting. This gives added emphasis to some problems of control.

CONTROL GROUPS The use of adequate control groups is especially important in most field experiments because the real life setting is both complex and changing, with many uncontrolled factors at work. Furthermore, we are frequently quite ignorant about these possibly confounding factors. An experiment on the effect of a supervisory training program upon employee morale, for example, might be confounded by such factors as a pay raise, changes in company personnel policies, etc. Such a study by Hariton showed that there were significant changes in the direction of the training goals even in the control groups (19). The trained foremen were matched with the control foremen on size of group, type of work, and level of morale, but the method of assigning subjects was by divisions of the company so that two divisions were placed in the experimental group and two in the control group. Later it was discovered

that the major determinant of the effectiveness of training was the human relations practices of the foreman's superior (the division manager) which, of course, was systematically different for the experimental and control groups. This is but one example of an organizational factor which can easily confound an experiment since it affects a whole group of subjects at the same time in the same way. Ideally the control group should be matched with the experimental group on *all possible confounding factors*, matching on the basis of easily measured variables such as size of group, age of subjects etc. will do no good unless these variables are systematically related to the hypothesis to be tested.

Solomon has shown (41) that the conventional design with a single control group is inadequate where the pre-experimental measures interact with the experimental treatment and influence its effectiveness. He conducted an illustrative field experiment on the effects of teaching spelling in a grammar school and demonstrated that the preliminary spelling test actually interfered with the training. The preliminary test resulted in smaller gains in the experimental group and in the conventional control group than were shown by a third matched control group which did not take the preliminary test but did receive the training. Solomon also mentions interaction effects and effects on variance in studies of attitude change. Canter has shown similar interaction effects in a study of human relations training (5). On *a priori* grounds one would suppose that certain measurement procedures such as open ended interviews, might produce even larger interaction effects. Accordingly, it seems desirable, where possible, to use an extended control group design in field experiments. This would involve the comparison of matched groups on post measures.

CONTROL THROUGH MEASUREMENT Where the experimenter is not able to standardize conditions, the effect of confounding factors may nevertheless be discovered through measurement. If he knows all the additional determinants of his dependent variable, the experimenter can measure them and determine by analysis whether any of them were confounded with his experimental variable. Thus, the field experiment will commonly require more measurement of the total situation and particularly of any uncontrolled changes occurring during the course of the experiment. If the experiment has been preceded by field studies in the same setting and

on the same problem it will be possible to know more of the possibly relevant factors which must be measured

REPLICATION Because of the difficulty of standardizing conditions and because conditions may vary over time it is desirable where possible to design field experiments that can be replicated in the same setting. For example, Bavelas ran a series of experiments on the effects of group decision on production in the Harwood Manufacturing Corporation (28). These experiments took place over a fairly long period of time and under somewhat varying conditions. The present author subsequently repeated some of the same experiments in the same setting at a later date (15). Such replication in the same setting is often difficult for group experiments because of an insufficient number of matched groups. To replicate an experiment in an organization setting there must be a number of parallel or matched units. Where the units or groups are not matched—i.e. where they are performing different functions or are differently organized—new and irrelevant variables may intrude themselves.

PRELIMINARY EXPERIMENTS Planning for replication of a field experiment even in the initial design has the further advantage of providing an opportunity to perfect the experimental manipulation and the controls. If replication is not planned it is all the more important to conduct preliminary experiments to work the bugs out of the experimental manipulations and to ensure that they will be sufficiently strong to produce measurable differences. The running of preliminary experiments has not in fact been done frequently enough. Experience from several field experiments would indicate the importance of providing ample opportunity to practice the necessary manipulations and at the same time to provide opportunity for testing measuring instruments.

STANDARDIZATION AND INSULATION So far we have discussed controlling conditions in a field experiment through the use of control groups and adequate measurement. We have described these methods first because they are the easiest rather than the best. Actually such control does not eliminate unwanted variation in conditions; it makes it possible merely to reduce the bad effects. There are two ways of experimental control which do eliminate or reduce such variation: standardization and what we shall call insulation.

Sometimes it is possible for the experimenter to standardize conditions in the sense of holding constant something which might otherwise vary in the field setting. In several studies of the effects of social variables (such as group decision) on production in a factory (30, Chap 9), it was possible to hold constant most of the variables affecting production—the type of machinery, the flow of materials, and the composition of the group (no transferring of employees into or out of the group was permitted). On the other hand, absenteeism and slight variations in the quality of materials could not be prevented. The field experimenter should attempt maximum standardization of such conditions.

Insulation is less clear as a form of control, but we shall use this term to refer to the elimination of certain conditions normally existing in the field setting. In the Hawthorne studies, for example, the Relay Assembly Test Room was insulated from the rest of the plant both physically and socially (37). There were walls separating the experimental group from the other employees, and the normal supervisory relation was eliminated. Evidently the conditions created were so different from the usual factory conditions that this experiment stands midway between a field experiment and a laboratory experiment. A much milder degree of insulation was used in the experiments on group decision, on pacing cards, and on production by instructing the supervisors not to say anything to their employees about level of production, speed of working, steadiness of working, etc (30, p 215). Thus insulation, by eliminating some variables, simplifies complex conditions so that the experiment becomes more like a laboratory experiment (see Chap 4).

In itself, such simplification is neither good nor bad, rather, it must be evaluated in terms of the same kind of considerations relevant for choosing how analytical to be (see p 105) or whether to use a laboratory experiment or a field experiment.

A special problem of insulation is the possibility that one experimental group may contaminate another experimental group or the control group. Communication among groups can lead to serious confounding. In an experiment on the effects of participation on production, for example, competition developed between two of the groups, thus confounding the effects of participation (9).

SAFETY FACTORS Because of the danger that the experimental manipulation may not be strong enough—in the sense that it does

not produce sufficiently large differences—it is desirable to design the experiment with certain “safety factors” so that some of the values of a field study may be obtained even if the experimental manipulation fails. For example, an experimental study of the effects of human relations training obtained measures of changes in both group structure and group process (4). Even though these variables showed little change that could be attributed to the training, there were nevertheless some interesting findings on the interrelations of process variables and structure variables.

SIZE OF UNITS AND LENGTH OF TIME. In general, it is probably much more effective to apply experimental treatments to units of fairly small size—individuals or small groups rather than large institutions. Dealing with smaller subgroups permits easier manipulation and provides an opportunity for preliminary experimentation and replication. The size of the experimental manipulation should also be reduced where possible by confining it to a relatively short period of time. The longer the period of time required, the greater the probability of unforeseen outside changes taking place. One has only to examine the changes over time in existing data on such variables as production or absenteeism in a factory, political and social attitudes, etc. to see that fairly large changes can take place in a few months. Thus the longer experiments are more likely to be confounded. Finally from the point of view of efficiency alone, it is desirable to reduce the time to a minimum.

CARRYING OUT A FIELD EXPERIMENT

It is not easy to describe the methods used and the skills required for carrying out the experimental manipulation and controls and for building and maintaining the necessary cooperative relationships, because these skills are not often described in detail in reports of research. To generalize about these skills is even more difficult for there has been little scientific study of them and they are not commonly taught in courses on research methods (6). Actually, the practical social managers and the professional consultants seem to be more skillful than the social scientists in bringing about a planned change in a social situation such as would be required to carry out an experimental design.

Accordingly, this section will draw heavily on the ideas of Kenneth Benne and the staff of the National Training Laboratory in Group Development concerning the skills of the practitioner (36). In discussing social action, they point out,

The behaviors of persons, groups, organizations and communities may be thought of as held in their accustomed grooves by an equilibrium of forces tending to move the level of functioning in one direction or another. These forces may be of various kinds and magnitudes—established status relations, laws, personal and group sanctions and standards, established competence in a given way of working, feelings of inadequacy, non perception of alternative ways of behaving, etc. Change, in the sense of an alteration in the level and way of functioning of a person or group, or an organization or community, occurs when this equilibrium of forces is disturbed. According to this view, change may be seen as the transition of a system of behavior from one equilibrium level to another. For example, a teacher may change from a manipulatory to a collaborative control relation to her group of students or vice versa, members of a group may change their participation pattern from one with a high percentage of individual centered roles to one with a low percentage of such roles, or vice versa; a community may change from a condition of competition between welfare agencies to a condition of coordinated cooperation, or vice versa, etc. . . . Planned change occurs when the forces holding the person or group or community at a given level, with respect to one or another phase of its life, can be assessed, when factors making for potential disequilibrium are understood, when a new possible and desirable state of affairs can be projected, and when the forces for effecting movement to this new equilibrium can be developed or manipulated to this end (36, p. 108).

This is also a good description of the complex process involved in the planned manipulation of a variable in a field experiment.

Our conception of how to bring about these changes is thus actually based on: (1) some theorizing about the process of change in terms of quasi stationary equilibria (30, Chap 9), (2) some conceptions of scientific methodology as an effective method of problem solving, (3) a methodological conception of democratic ethics, and

(4) some scientific research on the effectiveness of certain democratic procedures in bringing about planned changes. We shall not attempt to present the rationale for these foundations here but shall proceed directly to their implications for action. First we shall describe two experimental procedures and then attempt a few generalizations regarding the methods and the skills involved.

Examples of Experimental Treatments

A field experiment by Coch and French illustrates a method of manipulating the variable of *participation* and at the same time shows the effects of this variable in bringing about planned changes. (9) Since both authors had conducted earlier experiments in the same factory there was no problem of establishing the role of the researchers in the factory. Furthermore the experiment helped to solve one of the most difficult postwar problems of the factory. Because most of the products had to be changed as the plant reconverted to civilian production there were widespread technological changes affecting many employees. Earlier studies in the same factory had shown that such technological changes produced large decreases in production as well as strong resistance to the changes on the part of employees.

The experimenters wanted to test the hypothesis that the greater the participation on the part of employees in planning to meet the changed conditions the less would be their resistance to the change and their loss in production. This hypothesis was discussed with top management during the attempts to foresee the problems that would be caused by the changes and to diagnose their nature. Fortunately the management including the experimenters had experienced a considerable body of past research both on the effects of job changes on morale and productivity and on the effects of group decision in changing productivity. On the basis of the previous findings and of the research mindedness developed through previous participation in research it was easy to plan jointly not only a procedure for handling the technological changes by a method of participation but also an experimental design for systematically varying the degree of participation.

Three different degrees of participation were tried on matched groups. The no participation or control group was changed over

to the new job according to the usual factory procedure, which consisted of explaining, in a group meeting, why the changes were necessary, what the new job would consist of, and what the new piece rate would be. As usual, this procedure was carried out by the production manager and the time study engineer. The second treatment, "participation through representation," consisted of a group meeting, also conducted by the production manager, in which the need for a change was presented as dramatically as possible and the broader problem of cost reduction was shared with the group. This led to a discussion of the types of changes that should be made in the product and the suggestion that management should make work simplification studies, train several operators in the new methods, and set the piece work rates by time studies on these specially trained operators. The new job and the new piece rate would then be explained to all operators, with the special operators helping to train the others. These special representatives were then elected by the group. Later they met with the management to plan more about the new job, and they presented many good suggestions. These special operators became so involved that they referred to the new job as "our job" and "our rate." The third experimental treatment, "total participation," was much like the second except that all operators participated directly in planning the new job and the new rates rather than participating through representatives.

Although only a few hours were involved in the meetings which constituted the experimental manipulations, it was clear that they had strong and very different effects. Probably this was due partly to skillful leadership by the production manager and partly to the fact that he had an important role in the factory, which gave added meaning to his behavior. Although no measurements of the experimental treatments were made, the measured effects on production showed much higher productivity with greater participation.

This principle of participation is of fundamental importance in carrying out an experimental procedure. The production manager and other members of management were motivated to cooperate in the experiment because of outside economic pressures combined with recognized difficulties in handling such problems in the past. Certainly their past experiences with similar research on group decision contributed to both their understanding and their support.

Quite a different problem of experimental manipulation faced the research team in a study on changing attitudes in a housing project (14). A previous survey showed that members of the project had hostile attitudes toward one another, were ashamed of living in the project, and had little contact with either their neighbors or the surrounding townspeople. The experimenter wished to test the theory that stimulating contacts among project residents under favorable conditions would reduce the hostile attitudes, which in turn, would reduce the shame at living in the project, which, in turn, would cause the residents to initiate more social contacts with the town. The people in the project, however, felt no need for increased contact. In fact, on the basis of past failures, they were pessimistic about community activities. Furthermore, the expert community organizer on the research team had no clearly established role in the project. It was clear that she did not belong to the project, but it was not clear that she was related to the university research team.

The community organizer began with a meeting of residents in which she stimulated interest in a nursery school project, a recreation program for school age children, and adult education and recreation activities. Planning committees were formed, and a community wide meeting was held for further discussion of plans with resource experts. As progress was made in organizing these activities during the first month, and as more women became active, resistance to the activities developed, particularly on the part of the old established leaders. They apparently perceived the new activities as giving support to a possibly competing set of leaders. Eventually, a hostile rumor branding the activities as Communistic forced the suspension of all community activities for two weeks until the rumor could be combated by the presentation of detailed information about the sponsorship and purposes of the research program, by deliberate efforts to integrate the old leaders into the new activities, and by statements by the local project manager to the effect that the rumor had been demonstrated to be unfounded. Subsequently these and other community activities were started up again and continued over a total period of eight months. (The detailed record of this experimental procedure can be found in 14 pp. 28-44.) This experiment illustrates two errors which were later corrected but which probably could have been avoided in the first place if foresight had been

as good as hindsight. First, the role of the community organizer should have been structured more clearly for the residents. Secondly, the hostility of the community leaders should have been avoided by not appearing to support the competing leaders.

Cooperative Relations

We have already indicated (see p. 108) that cooperative relations must start in the planning phases of the project in terms of selecting a setting, a problem and a sponsoring organization in which the possibilities for mutual benefits from the research are maximal. Lippitt (31) summarizes the reasons for successful cooperation on a training workshop in intercultural relations as follows:

It seems probable in retrospect that this collaboration of social practitioners and social scientists was successfully initiated and carried through to fruition because

The social practitioners—

—felt keenly the inadequacy of present accomplishments in community improvement of intergroup relations

—were deeply sincere in their desires to achieve results rather than just to receive recognition for noticeable efforts

—believed in the potentialities for constructive change in the intergroup relations of conflicting groups in the community

—had a hunch that the methods for bringing about these changes might be more efficiently discovered by the application of scientific methodology than by the usual trial and error efforts of experience unguided by systematic fact finding

—were ready to have faith that this particular group of social scientists was sensitive to the requirements of an action job (i.e. running a workshop) as well as being adequate research technicians

—perceived that these social scientists had strong personal motivations to do something about intergroup relations in addition to a need for more scientific understanding

The social scientists—

—had a major interest in the scientific study of social change

—had arrived at the hypothesis that fruitful research on social change demanded an integrated service and study relationship to groups outside the walls of the laboratory

- had as a part of their research team personnel skilled in group action techniques as well as measurement methods

- had arrived at the conclusion that experimental methodology could be successfully applied to the study of the life of small and large groups in the democratic community

- had strong feelings that intercultural and interracial conflict besides being a challenging scientific problem was a priority social issue

This description fits Deutsch's theory of cooperation (11) Establishing a cooperative situation means creating a condition such that if one party achieves his purpose or goal, the other party at the same time achieves or is brought closer to his goal Conversely, it is important to avoid a competitive situation, in which locomotion of one party toward his goal will move the other party further from his goal Deutsch's research shows that, given a basic cooperative rather than a competitive situation, one can expect mutual helpfulness, better understanding of one another's communication improved coordination of efforts and other forms of behavior commonly considered cooperative (12)

The longer and more complex the experimental treatment, the more important it is to think in terms of *maintaining* the cooperative relationship by a constant awareness of one's own goals and motivations and how they are related to those of others involved in the experiment Probably most of the early resistances encountered in trying to carry out the experimental procedures in the housing study stemmed from the fact that psychologically there was no cooperative situation for the members of the housing project As the perception of a cooperative relation grew in more and more residents, these resistances seemed to decrease Thus we can see that establishing a cooperative relationship may be a gradual process which has to extend to more and more of the client organization during the course of the field experiment Two prescriptions are implied for the experimenter (1) try to clarify for all people concerned just what you are going to be doing when and with whom (2) provide enough time, enough contact and enough communication so that these people can have confidence in the experimenter, in the sense that they believe that he will do nothing to harm them

or to conflict with their interests and, indeed, that he has some positive concern for their welfare

The social scientist may be able to rely on the skill of an experienced practitioner, and cooperation with him will sometimes lead to the solution of the difficult problem of carrying out the experimental manipulation. This solution was successful in the experiment on participation because it was possible to obtain the intelligent cooperation of the production manager, who had the necessary skills to conduct the meetings successfully and a position of authority in the company. Many of the problems which have been attacked through field experiments have, in fact, been problems related to rather highly developed professional skills in such areas as human relations training, therapy, community organization, etc. For such problems there are already available highly skilled professional people who may be called upon to conduct the manipulation. The housing study is a case in point where a professional community organizer was employed as part of the research team. Ezriel points out that even so complex a skill of the psychoanalyst as the making of an interpretation in terms of unconscious impulses and fantasy determined reasons can be used to conduct replicable field experiments within the psychoanalytical session (13).

Executing the Experimental Treatment

Once the basic role relationships are worked out, there should be a *collaborative diagnosis* of the situation by the researcher and at least some part of the client organization. The purpose of this diagnosis is to assess the various factors that will be involved in executing the research design—the resistances that may be encountered, the dynamics of the situation in regard to the problem of bringing about a change, etc. The ways of going about such a collaborative diagnosis will vary tremendously, depending upon the problem to be studied, the setting, etc. In the example on participation, it was possible for all those involved in the diagnosis to rely on both an extensive practical experience in the situation and a considerable amount of relevant research data. Of course, all such relevant experience and information should be used, but whatever the source, a useful diagnosis will have to be formulated in theoretical

cal terms. Because it is necessary to predict reactions and to bring about planned changes, the diagnosis must go beyond the mere statement of facts or the labeling of things as 'good' or 'bad', it must move toward more causal thinking, for it is only by manipulating its causes that we can vary our dependent variable. Getting help from the practitioner is both a way for the scientist to increase his own diagnostic sensitivity and a method for maintaining the 'involvement' and interest of the practitioner (22).

The next step should be *joint planning*, based on an adequate diagnosis, of the actions that must be taken in order to manipulate the independent variable and to control other possibly confounding factors. To be most effective, this collaborative planning must go beyond the point of simply explaining to the client organization what is required in the research. It should also include joint decision making concerning the methods and techniques by which the design is to be carried out. It is desirable that the practitioners be enough involved in the research design so that they will fully understand not only what it is but some of the reasons lying behind it.

In effect, this usually means that the experimenter has to train some parts of the client organization in research methods. Many practitioners, for example, will have to learn more about scientific method before they are convinced of the necessity or the desirability of having a control group in a field experiment. A common reaction of management to a new selection device which looks promising is to want to employ it in selecting *all* new employees, whereas it will be necessary to continue selecting some employees on the old basis in order to evaluate the effectiveness of the new procedures. Here it is quite important to take time for the practitioner to learn why the control groups are necessary and how an increase in scientific understanding can contribute to his own practical objectives. Several experiments, including our example on the effects of participation, indicate that the more persons in the client organization who can be involved in the planning, the more widespread will be the support in carrying out the experiment.

In field experiments in social psychology one special problem of collaborative planning with the client organization frequently faces the experimenter—namely, the problem of secrecy. In order to carry out certain experiments, it may be necessary, at least for a

certain period of time, that all the subjects in the experiment, or possibly all of the people in the client organization, be prevented from learning fully the purpose of the experiment or some aspects of the procedure. Such secrecy is necessary if knowledge of the hypotheses would confound the results of the experiment. Of course, it is possible and desirable to give a full explanation at the conclusion of the experiment, but even where this is done the temporary secrecy can be a serious disturbance to good relations. In the housing study, for example, the experimenter was testing the hypothesis that improved interpersonal relationships within the housing project stimulated by the community activities would result in spontaneous initiation of communication between the housing project and the surrounding town (14). Accordingly, it was essential that the research team, while stimulating community activities within the project, should not also stimulate activities between the project and the townspeople. But it was also essential for adequate control that the subjects in the project not know about this reason for the experiment for fear they might initiate contact with the townspeople simply to please the experimenter. If this need for temporary secrecy is faced frankly in the early phases of planning for the experiment, and if it becomes a part of the training of the client in research methods, then it need not be a factor disturbing the cooperative relationship.

In planning techniques for the manipulation of variables, successful strategy and tactics for changing the variable must be based on a correct and adequate theory of change. For example, in the experiment on the relations between the housing project and the surrounding town, the manipulation of the degree of hostility of interpersonal relations within the project was based on the specific diagnosis that hostile interpersonal relations existed primarily because of autistic hostility and on the specific theory that these hostile relations could be changed by stimulating communication among members through any successful group activities. Of course, if these theories and the assumptions upon which they were based were incorrect, then the actual manipulation by stimulating activities in the project would not in fact change the variable of hostility of interpersonal relations. In the long run, therefore, improvements in our theory of change will increase our ability to plan successful manipulations.

Lewin (30, p. 193) has pointed out the importance of dealing with the total situation when manipulating a variable

To vary a social phenomenon experimentally the experimenter has to take hold of all essential factors even if he is not yet able to analyze them satisfactorily. A major omission or misjudgment on this point makes the experiment fail. In social research the experimenter has to take into consideration such factors as the personality of individual members, the group structure, ideology and cultural values, and economic factors. Group experimentation is a form of social management. To be successful it like social management has to take into account all of the various factors that happen to be important for the case in hand. Experimentation with groups will therefore lead to a natural integration of the social sciences, and it will force the social scientist to recognize as reality the totality of factors which determine group life.

For experiments which are more oriented to the testing of theoretical hypotheses, it is usually more difficult to use skilled practitioners. The manipulation will more frequently be one that has not been done before, or there will not be any large body of practical experience, thus it will more often be necessary for the experimenter to conduct the manipulation himself. In these field experiments, where the problem under investigation is actually some theoretical hypothesis, it is important that the experimental manipulation be valid in the same sense that a measurement should be valid—i.e., it should conform to the conceptual definitions involved in the hypotheses.

This is an additional difficulty in carrying out the experimental manipulation because it so greatly reduces the leeway for the experimenter. If, for example, one experiments on such a question as

What are the effects of role playing as compared to the effects of group discussion? the experimenter has great freedom in how he conducts these two manipulations, since they have such broad and various meanings. In contrast, Rosenberg's (38) experiment on role playing studied the more specific hypothesis that 'deeper emotional and attitudinal changes are produced the more the subject is deeply involved in the process of playing the role. In this case, it is important that the experimental manipulation actually con-

form to the conceptual definition of *degree of involvement in the role* rather than to some other similar independent variable. The methods used for creating three degrees of involvement were (1) the subjects actually participated in playing the key role in the sociodrama, (2) the subjects were instructed to identify with somebody else whom they observed playing the key role, (3) the subjects were instructed simply to observe what was going on in the group where the sociodrama was presented. Obviously, this procedure does vary the degree of involvement, but this variable is not defined clearly enough to permit accurate judgments as to how well the manipulations fit the definition. Probably other factors were varied simultaneously so that the experiment was to some extent confounded. For example, in one treatment, the attention of the subject was directed toward the key figure being observed, whereas in another treatment the attention of the subject was directed toward the group as a whole. One might say that direction of attention was systematically confounding these two degrees of involvement, thus reducing the validity of the manipulation.

Problems of Measurement

Especially where the field experiment is designed to study some theoretical hypothesis, it is essential for the researcher to measure the success of the manipulation used, for it is only by relating the results to some *known* procedure that knowledge can be advanced. Where one is dealing with complex experimental procedures, the measurements themselves may be quite complex, as in the case of the housing study. Even an extremely condensed reporting of the experimental manipulations used in this study required fifteen pages. In other cases, the manipulation has been measured quantitatively, as in the observational measurement of a variety of dimensions of the training process in a workshop (31, Chap. 11). In the more analytical experiments, the measurement of the success of the manipulation will be simpler. In Rosenberg's (38) experiment, the simplest measurement was to ask the subjects in a questionnaire whether they had identified with the key figure as instructed or observed the total group as instructed. The answers to this questionnaire showed that on the whole the instructions "took," but that there were some exceptions. When these latter subjects were recategorized

in the analysis of the data, it was found that the hypotheses were even more clearly supported than when the subjects were analyzed in accordance with the treatment to which they were assigned. Thus measuring the success of an experimental manipulation not only gives some indication of the degree to which it was successful but can also permit more valid manipulation through the elimination or reclassification of unsuccessful cases in subsequent analysis.

In a field experiment, most of the problems of social psychological measurement are the same as they are for other types of research (The reader is referred to Chapters 6, 11, and 12 of this book, which discuss measurement.) However, a word of caution should be added concerning one point. In most organizations, certain sorts of social psychological data already exist in the form of records used for administrative purposes. In industry, for example, one frequently finds records of productivity, of absenteeism, of turnover, and of other aspects of human behavior. It is tempting for the social psychologist to use all these data which are available with no expenditure of time, money, or effort on his part. Yet it seems to be widespread experience that such administrative records are a snare and a delusion for research purposes. First, since they are not designed for research purposes they do not always measure the exact variables demanded by one's hypotheses. Thus they tend to influence research away from a theoretical orientation. In the long run, this is probably a serious hindrance to the development of an adequate basic social science. Secondly, such records are usually very disappointing with respect to accuracy. For many purposes of the administrator, records need not be exact, and most of the personnel who collect or use such records are not trained in scientific standards of accuracy. Furthermore, the inaccuracies are frequently motivated and therefore concealed. One experimental procedure for example, resulted in a marked decrease in the variability of production which the researchers first interpreted as an indication of reduced tension (30, pp. 217-218). In a subsequent interview, however, we discovered that the employee had been turning in for the records only a part of her production on those days when her actual production was so high that she feared a cut in her piece rate. This word of caution is not to say that the researcher should always disregard records collected for administrative purposes, rather, he

should avoid being overoptimistic about their use and should use them only when he has an adequate check on their accuracy

Ethical Problems

Field experiments involve new and more difficult problems of professional ethics and some more difficult than those in laboratory experiments or field studies in which no changes are introduced by the experimenter. In a life setting, the experimental procedure is, in fact, social action, sometimes in a situation of social conflict, in which there are differences in values among the people involved (17). In such a situation it is especially important for the experimenter to be guided by a code of professional ethics. No generally accepted professional ethics, particularly for this branch of research, has yet been developed, although the American Psychological Association has been working on the formulation of such codes. Accordingly, this chapter will not attempt the difficult task, for the moment, each researcher must work out his own ethics.

It should be pointed out, however, that the experimental methods which have been recommended here are not ethically neutral. Maximizing the mutual benefit of the research to both the practitioner and the researcher, collaborative diagnosis, participation in planning the research, open dealings with client organizations, and educating the participant to understand the research—all are specifications of a democratic ethics. These methods of dealing with people are based on an explicit recognition of the ultimate value of each person and his right of self-determination.

A SUMMARY OF THE ROLE OF FIELD EXPERIMENTS

The practical advantages of a field experiment are most clear cut and simple. Anyone who wishes to take effective social action in any setting can improve upon the uncontrolled try-out of new methods by the application of more scientific experimental procedures. Through careful measurement, better theorizing, the use of control groups, and other aspects of improved experimental

design the practical problems of social action can be solved with greater certainty, with greater accuracy, and sometimes with greater efficiency than through common sense trial and error methods

The primary scientific advantage of the field experiment is that it permits a more unequivocal determination of causal relations it permits determining cause and effect where a field study would reveal only a correlation Secondly, we have seen that the field experiment is particularly appropriate for studying methods of social change, social processes and social influences Thirdly, the field experiment, since it deals with the total life situation, is well adapted for studying complex syndromes and social processes where the interrelationships among several more analytical variables are involved

But the very fact that it deals with the total life situation also leads to one of the major limitations of the field experiment—namely, it is not an appropriate method for studying with analytical precision more specific single hypotheses Finally, we must mention as a disadvantage of the method the difficulty in carrying it out because of the social skills required and the contacts necessary with settings which provide a good research opportunity Even where the skills and opportunities are maximal, many field experiments necessarily involve so long a span of time that they may be inefficient

The optimal scientific role for field experiments is in a program of research in combination with other methods They will be more successful if preceded by field studies which give a more extensive and exact knowledge of the setting and thus enable the experimenter to manipulate and control his variables more successfully The development of basic theory will widen the horizons for field experiments which test the range of application of generalizations arrived at in the laboratory More generally, to the degree that the field experiment attempts to test general hypotheses, it will make a contribution to science, otherwise it will have a more limited practical value

BIBLIOGRAPHY

- 1 Annis A D The relative effectiveness of cartoons and editorials as propaganda media *Psychol Bull*, 1939 36, 638
- 2 Back, K, Festinger, L, Hymovitch, B, Kelley, H, Schachter, S, and Thibaut, J The methodology of studying rumor transmission *Hum Relat*, 1950, 3, 307 312
- 3 Block, J, and Block, J An interpersonal experiment on reactions to authority *Hum Relat*, 1952, 5, 91 98
- 4 Bradford, L, and French, J R P, Jr (eds) The dynamics of the discussion group *J Soc Issues*, 1948, 4, No 2, 1 73
- 5 Canter, R R, Jr The use of extended control group designs in human relations studies *Psychol Bull*, 1951, 48, 340 347
- 6 Cartwright D Basic and applied social psychology *Phil Sci*, 1949, 16, 198 208
- 7 Chein, I, Cook S, and Harding, J The use of research in social therapy *Hum Relat*, 1948, 1, 497 511
- 8 ——— The field of action research *Amer. Psychologist*, 1948 3, 43 50
- 9 Coch, L, and French, J R P, Jr Overcoming resistance to change *Hum Relat*, 1948, 1, 512 532
- 10 Coffey, H, Freedman M, Leary, T, and Ossorio, A Community service and social research—group psychotherapy in a church program *J Soc Issues*, 1950, 6, No 1, 1 65
- 11 Deutsch, M A theory of cooperation and competition *Hum Relat*, 1949, 2, 129 152
- 12 ——— An experimental study of the effects of cooperation and competition upon group process *Hum Relat*, 1949, 2, 199 232
- 13 Eziel, H The scientific testing of psycho analytic findings and theory *Brit J Med Psychol*, 1951, 24, 30 34
- 14 Festinger, L., and Kelley, H *Changing attitudes through social contact* Ann Arbor Research Center for Group Dynamics, Institute for Social Research, Univ of Michigan, 1951
- 15 French, J R P, Jr Field experiments changing group productivity In Miller, J G (ed) *Experiments in social process* New York McGraw Hill, 1950, pp 79 96

- 16 French J R P Jr, Kornhauser, A, and Marrow, A J (eds) Conflict and cooperation in industry *J Soc Issues*, 1946 2, No 1, 1-55
- 17 ———, and Zander, A *The group dynamics approach to psychological studies of labor management relations* A paper presented at the meeting of the American Psychological Association, Denver, 1949
- 18 Greenwood E *Experimental sociology, a study in method* New York Kings Crown 1945
- 19 Hariton T *Conditions influencing the effects of training foremen in new human relations principles* Ph D thesis, Univ of Michigan, 1951
- 20 Hartmann, G W A field experiment on the comparative effectiveness of 'emotional' and 'rational' political leaflets in determining election results *J Abnorm Soc Psychol*, 1936 31, 99-114
- 21 Jackson, J M An experimental investigation of leadership in a work situation (To be published in *Hum Relat*)
- 22 Jacobson E, Kahn R L, Mann F C, and Morse, N C Research in functioning organizations *J Soc Issues*, 1951, 7, No 3 64-71
- 23 Jahoda, M, Deutsch, M, and Cook S W *Research methods in social relations Part I Basic processes* New York Dryden Press, 1951
- 24 Jaques, E *The changing culture of a factory* London Tavistock Publications 1951
- 25 ——— (ed) Social diagnosis and social therapy *J Soc Issues*, 1947, 3, No 2, 1-68
- 26 Jenkins, D H, and Lippitt, R *The interpersonal perceptions of teachers, students, and parents* Washington National Education Assoc, 1951
- 27 Krech, D (ed) Action and research—a challenge *J Soc Issues*, 1946, 2, No 4, 1-79
- 28 Lewin, K Group decision and social change In Newcomb, T, and Hartley, E *Readings in social psychology* New York Holt, 1947, pp 330-344
- 29 ——— *Resolving social conflicts* New York Harper, 1948
- 30 ——— *Field theory in social science* New York Harper, 1951
- 31 Lippitt, R *Training in community relations* New York Harper, 1949
- 32 ——— The strategy of sociopsychological research In Miller, J G (ed) *Experiments in social process* New York McGraw Hill, 1950 pp 17-30
- 33 Mann, F Changing superior subordinate relationships *J Soc Issues*, 1951, 7, No 3, 56-63
- 34 Morse N C An experimental study in an industrial organization In

- Guetzkow, H (ed) *Groups, leadership and men* Pittsburgh Carnegie Press, 1951, pp 96-99
- 35 ———, Reimer, E., and Tannenbaum A S Regulation and control in hierarchical organizations *J Soc Issues*, 1951, 7, No 3 41-48
- 36 National Training Laboratory in Group Development *Bulletin No 3, Report of the second summer laboratory session* Washington National Education Assoc., 1948
- 37 Roethlisberger, F J, and Dickson, W J *Management and the worker* Cambridge Harvard Univ Press, 1939
- 38 Rosenberg, P P *An experimental analysis of psychodrama* Ph.D thesis, Harvard Univ, 1951
- 39 Selltitz, C., and Cook, S W Can research in social science be both socially useful and scientifically meaningful? *Amer Sociol Rev*, 1948, 13, 454-459
- 40 Snyder, R *An experimental study of the influence of leaders in small work groups* Ph.D thesis Mass Inst of Tech, 1953
- 41 Solomon, R L An extension of control group design *Psychol Bull*, 1949, 46, 137-150
- 42 Star, S A., and Hughes H M Report on an educational campaign the Cincinnati Plan for the United Nations *Amer J Sociol*, 1950, 55, No 4, 389-400
- 43 Van Zelst, R H Validation of a sociometric regrouping procedure *J Abnorm Soc Psychol*, 1952 47, No 2 Supplement, 299-301
- 44 Worthy, J C *Attitude surveys as a tool of management* American Management Association, General Management Series No 145

Laboratory Experiments

Leon Festinger

Empirical science in general has as its major objective the understanding or control of phenomena as they occur in the real world. Nevertheless, laboratory experimentation generally plays a significant part in the development of a science. It is important to have some understanding of why this should be true and of the exact function which laboratory experimentation should have in relation to the science as a whole.

We shall, consequently, attempt to clarify two aspects of laboratory experimentation—namely, what a laboratory experiment is and how the results of such experiments can be applied to the real world. It would be relatively easy to discuss the role of laboratory experimentation by means of examples from the physical sciences, but we shall attempt, rather, to illustrate the points to be made by examples from the problem area of social psychology. Although by doing this we may not be able to make our point as clearly as would otherwise be the case, we hope that the discussion will be more meaningful and carry more weight if it is entirely oriented toward the field which is now under consideration.

THE NATURE OF LABORATORY EXPERIMENTATION

What Constitutes a Laboratory Experiment in Social Psychology?

A laboratory experiment may be defined as one in which the investigator creates a situation with the exact conditions he wants to have and in which he controls some, and manipulates other, variables. He is then able to observe and measure the effect of the manipulation of the independent variables on the dependent variables in a situation in which the operation of other relevant factors is held to a minimum. Such a definition is, however, a great oversimplification. Given the techniques of experimentation today available, an investigator can at best achieve only a rough approximation of the degree of precision implied by the definition. As better techniques are developed, more control over laboratory experiments will, of course, be possible. At present, however, we must include under the term "laboratory experiment" a wide range of studies with varying degrees of control and precision.

We shall attempt, largely by means of examples, to distinguish between what might properly be called "field experiments" and "laboratory experiments." In many cases, of course, the distinction is clear and easy to make, in other cases it is difficult to maintain. In general, we shall be guided by the two parts of our definition: whether or not there was an attempt to create a specially suited situation, and the degree of precision in the control and manipulation of variables.

It would seem clear that experiments in industry such as have been described in the preceding chapter should not be called laboratory experiments. There is little or no attempt to set up special conditions. Typically, the situation is accepted as it is found and some manipulation is imposed. The manipulation of the independent variable is usually a simultaneous manipulation of a set of factors. The degree of control obtained in these experiments is usually not sufficient to guarantee that the effects observed are unequivocally related to the manipulation of the independent variable.

Let us compare such field experiments with the Lewin, Lippitt, and White study (21) on autocratic and democratic atmospheres. This was a relatively early experiment in social psychology and is perhaps close to the boundary between laboratory and field experiments. In this study a number of boys' clubs were set up for the express purpose of performing the experiment. There was no real life situation which was taken as given. Rather, a special set of circumstances was created because it was felt that the situation thus achieved would be an appropriate one for the study of the variables in which the experimenters were interested. In this sense it should properly be called a laboratory experiment, although its precision is perhaps not very much greater than the precision of an experiment in industry, such as the one reported by Coch and French (7).

In the Lewin, Lippitt, and White experiment, the manipulation of the independent variables consisted in having one leader of a boys club behave in a certain prescribed manner as compared to another leader of another club who behaved quite differently. These two sets of behavior, which produced measurable differences in the behavior of the club members, were complex and differed in many dimensions. The experimenters were undoubtedly not clear about all aspects of the differences created. Thus, rather than isolating and precisely manipulating a single variable or small set of variables, the experimenters attempted a large and complex manipulation. There was also little attempt at control in setting up the clubs. In terms of the control achieved and the degree of refinement in manipulation of the independent variables, this study is probably indistinguishable from most field experiments.

We shall now consider as an example of a laboratory experiment with a relatively high degree of control and precision, an experiment by Festinger (10) on voting behavior. In this experiment an attempt was made to vary a single factor—namely, whether or not the subjects knew the religious affiliation of the other members of the group. Groups were set up for the express purpose of the experiment, with care taken to ensure that every member of the group was initially a stranger to every other member. Exactly comparable conditions were created for each group. The nominees for whom subjects voted were always paid participants whose behavior was standardized. These same paid participants identified

themselves as having different religions in the different experimental groups, thus controlling for a wide variety of personality factors and first impressions

In such an experiment we can be more certain than we can in a field experiment that the results obtained are due directly to the variable manipulated by the experimenter. It is probable that a variable such as 'whether or not the subjects know the religious affiliation of the other members' is still not a fine or precise factor; it is probably once more, a cluster of factors. A laboratory experiment should, however, attempt to refine the manipulations as much as the present state of knowledge permits. One of the marks of progress in a science is the extent to which such laboratory manipulation can be refined and specified.

There is frequently a tendency in social psychology to criticize laboratory experiments because of their 'artificiality'. A word must be said about this criticism, because it probably stems from an inaccurate understanding of the purposes of a laboratory experiment. A laboratory experiment need not, and should not, be an attempt to duplicate a real life situation. If one wanted to study something in a real life situation, it would be rather foolish to go to the trouble of setting up a laboratory experiment duplicating the real life condition. Why not simply go directly to the real life situation and study it? The laboratory experiment should be an attempt to create a situation in which the operation of variables will be clearly seen under special identified and defined conditions. It matters not whether such a situation would ever be encountered in real life. In most laboratory experiments such a situation would certainly *never* be encountered in real life. In the laboratory, however, we can find out exactly how a certain variable affects behavior or attitudes under special, or pure, conditions.

This is certainly not the end of the task. One must also find out how these variables interact with other variables. The possibility of application to a real life situation arises when one knows enough about these relationships to be able to make predictions concerning a real life situation after measurement and diagnosis of the state of affairs there.

Let us compare such field experiments with the Lewin, Lippitt, and White study (21) on autocratic and democratic atmospheres. This was a relatively early experiment in social psychology and is perhaps close to the boundary between laboratory and field experiments. In this study a number of boys' clubs were set up for the express purpose of performing the experiment. There was no real life situation which was taken as given. Rather, a special set of circumstances was created because it was felt that the situation thus achieved would be an appropriate one for the study of the variables in which the experimenters were interested. In this sense it should properly be called a laboratory experiment, although its precision is perhaps not very much greater than the precision of an experiment in industry, such as the one reported by Coch and French (7).

In the Lewin, Lippitt, and White experiment, the manipulation of the independent variables consisted in having one leader of a boys' club behave in a certain prescribed manner as compared to another leader of another club who behaved quite differently. These two sets of behavior, which produced measurable differences in the behavior of the club members, were complex and differed in many dimensions. The experimenters were undoubtedly not clear about all aspects of the differences created. Thus, rather than isolating and precisely manipulating a single variable or small set of variables, the experimenters attempted a large and complex manipulation. There was also little attempt at control in setting up the clubs. In terms of the control achieved and the degree of refinement in manipulation of the independent variables, this study is probably indistinguishable from most field experiments.

We shall now consider, as an example of a laboratory experiment with a relatively high degree of control and precision, an experiment by Festinger (10) on voting behavior. In this experiment an attempt was made to vary a single factor—namely, whether or not the subjects knew the religious affiliation of the other members of the group. Groups were set up for the express purpose of the experiment, with care taken to ensure that every member of the group was initially a stranger to every other member. Exactly comparable conditions were created for each group. The nominees for whom subjects voted were always paid participants whose behavior was standardized. These same paid participants identified

effect of friendship existed in the absence of differential amounts of contact. It would enable one to accept or reject the third hypothesis stated above. In other groups one could experimentally vary the accessibility of other members for communication to obtain evidence as to whether the friendship represented a decrement in restraint against communication or whether there were actual pressures to communicate in the specific direction of friends.

Such an experiment would undoubtedly be difficult to set up, but, since the major body of this chapter will be devoted to the discussion of how to perform such experiments and how to produce the desired conditions, we shall not, at the moment, go into the details of how it might be done. Let it suffice now to say that in the laboratory, by setting up an artificial situation, we should be able to verify, elaborate, and refine our knowledge so as to increase our understanding of important processes in social life. It should be stressed again, however, that the problem of application of the results of such laboratory experiments to the real life situation is not solved by a simple extension of the result. Such application requires additional experimentation and study. It is undoubtedly important that the results of laboratory experiments be tested out in real life situations. Unless this is done the danger of 'running dry' or 'hitting a dead end' is always present. A continuous interplay between laboratory experiments and studies of real life situations should provide proper perspective, for the results obtained should continually supply new hypotheses for building the theoretical structure and should represent progress in the solution of the problems of application and generalization.

Difficulties of Performing Laboratory Experiments

Laboratory experiments, however, do not represent an easy road to the collection of data for the resolution of theoretical problems. In social psychology they are typically difficult to do, and many dangers are present in their execution. It is extremely difficult to create in the laboratory forces strong enough for results to be measurable. In the most excellently done laboratory experiment, the strength to which different variables can be produced is extremely weak compared to the strength with which these variables exist and operate in real life situations. One is able to obtain

The Relationship Between Laboratory Experimentation and the Study of Real life Situations

In the conducting of research, there should be an active interrelation between laboratory experimentation and the study of real life situations. It is relatively rare in social psychology that hypotheses, hunches, and recognition of important variables emerge initially from the laboratory, most often they arise in either the formal or the informal study of real life situations. In studying real life situations, we are forced to deal with the factors and variables as they exist in all their complexity. Because of this complexity and lack of control, it is rather rare that definitive conclusions and unequivocal interpretations are reached in such studies, but frequently new variables and new hypotheses are brought to our attention. One can take these suggestions, hypotheses, and hunches and use laboratory experimentation to verify, elaborate, and make more secure the theoretical basis for the empirical results which have been obtained.

In the laboratory experiment, sufficient control can be achieved to obtain definitive answers, and systematic variation of different factors is possible. As a result of this greater control, precision, and manipulability, conclusive answers can be obtained and relatively precise and subtle theoretical points can be tested. For example, in a study of the spread of a rumor in a community (11), it was found that the more friends people had, the more likely they were to have heard the rumor. This finding may suggest the hypothesis that friendship reduces restraints against communication of various types of content, or it may suggest the hypothesis that the existence of a friendship makes for an active pressure to communicate, or it may suggest the hypothesis that those who have more friends see more people and spend more time with these people and consequently are more likely to have an opportunity to hear the rumor. In a laboratory experiment it would be possible to set up a situation in which one could, with a high degree of rigor, collect data which would enable one to choose among these possible interpretations. One could, for example, form groups of strangers and friends mixed together in which the amount of contact among members and the opportunity for communication among them were experimentally held constant. The results would enable one to say whether the

effect of friendship existed in the absence of differential amounts of contact. It would enable one to accept or reject the third hypothesis stated above. In other groups one could experimentally vary the accessibility of other members for communication to obtain evidence as to whether the friendship represented a decrement in restraint against communication or whether there were actual pressures to communicate in the specific direction of friends.

Such an experiment would undoubtedly be difficult to set up but, since the major body of this chapter will be devoted to the discussion of how to perform such experiments and how to produce the desired conditions, we shall not, at the moment, go into the details of how it might be done. Let it suffice now to say that in the laboratory, by setting up an artificial situation, we should be able to verify, elaborate, and refine our knowledge so as to increase our understanding of important processes in social life. It should be stressed again, however, that the problem of application of the results of such laboratory experiments to the real life situation is not solved by a simple extension of the result. Such application requires additional experimentation and study. It is undoubtedly important that the results of laboratory experiments be tested out in real life situations. Unless this is done the danger of 'running dry' or 'hitting a dead end' is always present. A continuous interplay between laboratory experiments and studies of real life situations should provide proper perspective, for the results obtained should continually supply new hypotheses for building the theoretical structure and should represent progress in the solution of the problems of application and generalization.

Difficulties of Performing Laboratory Experiments

Laboratory experiments, however, do not represent an easy road to the collection of data for the resolution of theoretical problems. In social psychology they are typically difficult to do, and many dangers are present in their execution. It is extremely difficult to create in the laboratory forces strong enough for results to be measurable. In the most excellently done laboratory experiment, the strength to which different variables can be produced is extremely weak compared to the strength with which these variables exist and operate in real life situations. One is able to obtain

results and to see clearly how these variables operate, in spite of this weakness because of the increased control one has in the laboratory situation. But it is always possible, even probable, that the factors will be so weak that no differences between conditions experimentally created are apparent in spite of the increased control. Thus, in the setting up of a laboratory experiment, especial care must be taken to make the variables as strong as one possibly can. Unfortunately, one can determine whether or not one has succeeded only after the experiment is over. An exception to this generalization about the weakness of laboratory manipulation can be seen in Asch's use of the announced perceptions of group members (2). This involved, however, the use of seven confederates for a single experimental subject.

Related to the problem of the strength of forces in the laboratory situation is the difficulty of manipulating several variables simultaneously. In the complex field of research with which we are here concerned, it is frequently theoretically important to see the effect of the simultaneous operation of two or more variables. Unfortunately, however, the more variables the experimenter attempts to manipulate, the lower will be the strength of each variable. This is especially true if the manipulation of the variable is to be done by means of verbal instructions to the subjects. The result of this is at least at the present stage of technical development, that the number of variables which it is possible to manipulate simultaneously in the laboratory is relatively restricted. This will undoubtedly become less true as more powerful techniques of manipulating variables in the laboratory are developed.

These difficulties have an important implication for the conclusions one can draw from the results of laboratory experiments. As in any study, it is possible that the experimenter is dealing with entirely irrelevant variables—that is, there may actually be no relationship among the variables that are being studied. Such a condition would result in negative results—that is, no differences between experimental and control groups. However, we should also find a lack of differences between experimental and control conditions if our experimental manipulations were not sufficiently strong to reveal measurable differences even though such differences really exist. Thus, negative results from a laboratory experiment can mean very little indeed. If we obtain positive results—that is,

demonstrably significant differences among conditions—we can be relatively certain concerning our interpretation and conclusion from the experiment. If, however, no differences emerge, we can generally reach no definitive conclusion unless we are quite certain that the manipulation of variables in the experiment was done successfully and adequately. At the present stage of technical development, it is seldom that we can be certain, in the absence of positive results, that our manipulations were adequate. Undoubtedly, as more and more experiments are done, good evidence will become available for believing that a certain manipulation is an adequate one, and then negative results can be interpreted as demonstrating no relationship. At the present time, however, it is all too easy to set up a laboratory experiment which because of the ineffective manipulation of variables, will show no differences among conditions. It should be stressed again that at the present stage of technical development, negative results perhaps reveal only the fact that the experiment was not set up carefully and that the experimenter's attempted manipulation of the variables was ineffective.

Keeping in mind these difficulties and the relationship which must exist between laboratory and field investigation we shall now proceed to a more detailed examination of how laboratory experiments can be performed.

THE DESIGN OF LABORATORY EXPERIMENTS

The first and foremost requirement for a successful laboratory experiment is that the problem be stated in experimental terms. This means essentially that there must be a high degree of specificity and clarity in the statement of the problem and in the definition of the variables involved. The foregoing implies that before one can successfully do a laboratory experiment one must already know quite a bit about the phenomena one is investigating.

The process of specifying and clarifying the statement of a problem so that it is amenable to experimental treatment is by no means a simple or easy one. Let us take an example to illustrate the kinds of problems which confront the experimenter at this stage. In a field study of transmission of a rumor in an organiza-

tion (4), it was observed that communication tended to be directed upward in the organizational hierarchy. This result was explained as depending upon forces acting on members to move upward in the organization, i.e., the upward communication represented substitute movement on the part of the members.

Kelley (19) set out to perform a laboratory experiment to test this hypothesis more thoroughly. At this point the statement of his problem might have been 'What direction does communication tend to take in a structured hierarchy?' This statement, however, is still much too general and vague for the purposes of an experiment. An attempt to think in terms of setting up an experiment makes it immediately clear that one must answer questions such as

'What exactly is a hierarchy?' and 'Exactly what kinds of communication are we talking about?' There are many aspects to what is customarily thought of as a hierarchical structure. Do superior levels in the hierarchy have power over subordinate levels and if so, what kinds of power? Is each successive level upward in the hierarchy characterized by increased attractiveness of the work, or increased freedom of choice of what work to do, or increased importance of the work? For the purpose of setting up a laboratory experiment the theory involved and the definition of hierarchy must be made more specific. Kelley chose to establish a hierarchy in the laboratory on the basis of the perceived importance of the job to the subjects, holding the actual attractiveness of the job and the exact work that was done constant for both levels in the hierarchy.

Let us now consider the question of what kind of communication would be expected to go upward in such a hierarchy. It was clear that a distinction had to be made between work oriented communication, communication of criticism, communication of information, and communication which was irrelevant to the task. It was largely in the last category of communication content that the effect of substitute movement would be expected to appear. Consequently, the experiment was set up to allow and in fact to encourage communication of irrelevant content. The final problem in Kelley's experiment was phrased as 'What is the direction of irrelevant communication content in a hierarchy based upon perceived differential importance of the task?' This statement was specific enough to permit the design of the actual experiment. This

process of clarifying the objectives of the experiment takes considerable time, although it may not take long to describe after it has once been done

The difficulties of designing a laboratory experiment are by no means overcome when the problem has been specifically defined. There remain the major tasks of inventing measurement devices and techniques for manipulation of variables which will clearly measure and manipulate the variables which have been defined in the statement of the problem. No matter how specifically and clearly the concepts are defined in the statement of the problem, the laboratory experiment cannot be successful unless the measurement and the manipulation of variables actually relate to these defined concepts.

Thus, for example, in the Kelley (19) experiment mentioned above, it was necessary to develop techniques for producing a hierarchy as defined, while other variables, such as the type of work done, power, and attractiveness, would be controlled. The situation created had to be one in which irrelevant communication would occur. Adequate techniques for measuring the amount and direction of communication had to be developed. In the experiment, a two level hierarchy was established. Each level did exactly the same kind of work, although each was under the impression that the other level was doing something different. High and low hierarchical perceptions were encouraged by the instructions to the subjects. One subgroup was told that its own job was the important one, the other subgroup was told that the job of the other level was the more important. Communication of irrelevant material was encouraged by having all communication carried on in writing and by injecting into the communication stream prepared fictitious notes which were irrelevant in their content, thus encouraging subjects to do such writing themselves. All notes were collected and kept, and thus analysis of the content of the communication its direction, and amount was possible.

It is rarely safe to assume beforehand that the operations used to manipulate variables will be successful and will tie in directly with the concept the experimenter has in mind. It is a worthwhile precaution to check on the success of the experimental manipulations. In the experiment by Kelley, the subjects were asked a number

of questions after the session was over to determine whether or not the manipulation of status in the hierarchy had been successful. It was found that, in terms of their reported perception of status and their desire to be in the other role, the manipulation had created a difference between the two levels. This difference was a relatively small one, however. Small differences in the results could be directly attributed to the small difference in perceived status. When the difference in perceived status was made larger by selecting out those subjects for whom the experimental manipulations were clearly successful, the results become much clearer and more conclusive. If there had been no check on the success of the experimental manipulation, such analysis would have been impossible. It would also have been impossible to attribute unequivocally the inconclusiveness in the results to the relative inadequacy of the experimental manipulation.

The problem of the adequacy of the manipulation of variables may be dealt with in part by preliminary studies. In almost any laboratory experiment, the initial design will have certain inadequacies which will become clear after a few trial experiments. Such preliminary runs are also important to provide practice for the investigator so that his behavior and his instructions become standardized by the time the regular experiments start.

THE EXECUTION OF LABORATORY EXPERIMENTS

Techniques of measurement, manipulation, or control of variables can be introduced at almost any stage in the process of a laboratory experiment. We shall attempt, in the following pages, to cover in detail most of the techniques which have been used fruitfully and to give examples of their successful use.

Decisions about Subjects for the Experiment

Decisions about the kinds of persons to be used as subjects, how they are to be recruited, and what they are to be led to expect before they come to the experiment provide important opportunities for the manipulation of variables.

Controlling the Composition of the Group

It is possible to arrange the composition of the group so as to control the number of friends in each group or to select subjects to ensure that all of the members of a group are strangers to one another at the beginning of the experiment. The decision concerning the composition of the group depends of course upon the purpose of the experiment and on the variables upon which the experimenter desires to focus his investigation. We shall give some examples of the introduction of an experimental control or manipulation at this stage of the procedure.

The experiment by Festinger (10) previously referred to had as its objective the determination of whether knowledge of religious affiliation in a mixed Catholic Jewish group would affect the attitudes of members toward one another. It was assumed that these attitudes would be reflected by their votes in elections for officers of a club. It was decided to have groups meet in the laboratory and elect officers of a club into which they formed themselves. Half of the elections were to take place while no one in the group knew the religious affiliation of any one else, the other half of the elections were to take place after the religious affiliation of each member was publicly announced. It was obviously essential, for this procedure to be successful, that none of the six members of any group know one another. Contact was made with nine colleges in the Boston area and permission to recruit volunteers in each college was obtained. Experimental sessions were then scheduled so that in each group only one person from any one college was present. Thus, when the group met, the six members each came from a different college in the area and the chances of their knowing each other were quite low. In spite of all these precautions however, one out of 13 groups had to be eliminated because two of the members did know each other, having gone to high school together. In the other 12 groups, all the members were complete strangers to one another.

Schachter (26), in an experiment designed to investigate the relationships between difference of opinion and rejection also wanted his groups composed of strangers to minimize the effects of past history, such as established preferences or aversions, among members. Having strangers was important because he was partic

of questions after the session was over to determine whether or not the manipulation of status in the hierarchy had been successful. It was found that, in terms of their reported perception of status and their desire to be in the other role, the manipulation had created a difference between the two levels. This difference was a relatively small one, however. Small differences in the results could be directly attributed to the small difference in perceived status. When the difference in perceived status was made larger by selecting out those subjects for whom the experimental manipulations were clearly successful, the results become much clearer and more conclusive. If there had been no check on the success of the experimental manipulation, such analysis would have been impossible. It would also have been impossible to attribute unequivocally the inconclusiveness in the results to the relative inadequacy of the experimental manipulation.

The problem of the adequacy of the manipulation of variables may be dealt with in part by preliminary studies. In almost any laboratory experiment, the initial design will have certain inadequacies which will become clear after a few trial experiments. Such preliminary runs are also important to provide practice for the investigator so that his behavior and his instructions become standardized by the time the regular experiments start.

THE EXECUTION OF LABORATORY EXPERIMENTS

Techniques of measurement, manipulation, or control of variables can be introduced at almost any stage in the process of a laboratory experiment. We shall attempt, in the following pages, to cover in detail most of the techniques which have been used fruitfully and to give examples of their successful use.

Decisions about Subjects for the Experiment

Decisions about the kinds of persons to be used as subjects, how they are to be recruited, and what they are to be led to expect before they come to the experiment provide important opportunities for the manipulation of variables.

it may be more difficult to manipulate variables adequately. On the other hand, if the experimenter plans on more than one meeting per group, he must expect that a certain percentage of subjects will not return after the first meeting.

Designs which require the group to meet several times encounter another difficulty. Many uncontrolled factors may be introduced since the subjects may contact one another outside the experiment and, in this way, materially change the situation between experimental meetings. The decision as to which of these two types of experimental designs to employ depends, again, upon the objectives of the experiment and on how these objectives can best be accomplished. A number of examples of each kind of experiment will be given to illustrate the advantages and difficulties.

Deutsch (9), in his study of the effects of competitive and cooperative situations on group problem solving, felt that the full effects of the experimental variables would reveal themselves only if the group would have considerable experience working together under the prescribed conditions. He decided on six successive meetings of each group and, to accomplish this, persuaded the instructor of a course to give students credit for participating in his experiment. Under these conditions most subjects attended all six sessions. Such an arrangement is not usually possible, but it is generally necessary to have some means of ensuring that subjects will return when the group is to meet several times.

Schachter's (26) experiment on rejection of deviates used one meeting of each group. It was necessary, however, for the subjects to be under the impression that they were to continue to meet once a week for a considerable period of time. The experimenter recruited subjects by telling them about clubs that were being formed and giving them the opportunity to join one of the clubs. Subjects were told that by joining they were committing themselves to attend the first meeting. After the first meeting they would be able to decide for themselves whether or not they wanted to continue.

In an experiment on strength of attraction to groups, Libo (22) used the number of meetings which subjects attended as one of the major measures of the strength of their attraction to the group. He, too, gave subjects an opportunity to volunteer to join clubs which were to continue to meet every week. Subjects could decide,

ularly concerned with the effect of the experimental condition upon acceptance and rejection. He recruited volunteers from courses which were divided into small recitation sections. By scheduling, in any one group, only one person from any one recitation section, he was fairly successful in eliminating prior acquaintanceship.

In both the examples above, having strangers compose the group was a technique used to exercise additional control over the experimental situation. In experiments on the effects of discussion on opinions about matters of fact, Jenness (18) controlled the range of difference of opinion in the group by the assignment of subjects to given groups on the basis of their original estimates of the facts in question. French (16), in an experiment on the effects of frustration and fear, used the composition of the group as a means of manipulating a variable. He was concerned with the differential effects of frustration and fear upon organized and unorganized groups. For his unorganized groups he used subjects recruited at Harvard University who met together as a group for the first time in his laboratory. For his organized groups he used club members who had a long history of working together and engaging in activities as a group. The members of each organized group came to the laboratory together. This type of manipulation is, of course, a gross one, since an organized group is different in many ways from an unorganized one. The same type of manipulation of the composition of a group can, however, be used in any number of ways to produce fine or gross differences among conditions. Some of the earliest experiments with groups, for example, employed as their major variable the presence or absence of other persons (1). Whether the person worked alone or in a group of people or before an audience was found to affect his performance (8).

Duration of the Group's Existence

Before recruiting subjects, it is necessary to decide whether the experiment will be conducted in one meeting or whether the group will be required to continue for several sessions. Each of these procedures has advantages and disadvantages. If the experiment is to be performed in only one meeting, it is generally easier to obtain volunteers. If the experimenter is restricted to one session, however,

joining that specific club (low attraction to the group)¹ This manipulation of attraction to the group was also shown to be successful by the results and by answers which subjects made to questionnaires after the experiment

Size of the Experimental Groups

No matter what techniques the experimenter employs there will always be some subjects who, after having agreed to be at the laboratory at a certain time, will not appear. They may have forgotten, they may have changed their minds or something may have happened which made it impossible for them to attend. In any event, the problem for the experimenter is the same. In designing a laboratory experiment in which human subjects are to be used, it is well either to design the experiment so that it may be conducted with a variable number of subjects or to make some provision to ensure the proper number of persons in each group. It is generally most desirable to allow for variation in the number of subjects. Thus for example, an experiment may be designed so that it can be conducted with either five, six, or seven members in the group. If seven persons are then scheduled for each meeting, and if sufficient precautions are taken,² very few groups will be lost.

When a design requires a constant number of subjects in each group, there are a number of techniques to ensure the presence of the proper number. Festinger (10), in his experiment on the effects of knowledge of religious affiliation, felt it necessary to keep the size of the groups exactly constant at six subjects per group. Three of these were to be Jewish and three Catholic. This was essential because of the desire to have the group evenly divided between the two religions. Leeway in the number of subjects in each group would have produced deviations from an even division which might have introduced additional complexities. Before each experiment

¹ This is not strictly an experimental manipulation of a variable. Rather it represents selection of subjects on the basis of some measure in order to create contrasting conditions.

² There are many factors which will affect the proportion of subjects who having volunteered actually come to the experiment. If for example volunteers are recruited from university classes the more pressure applied upon them to participate the lower the proportion of subjects who appear when scheduled (27).

after the first meeting, whether they wanted to continue their membership. Little pressure was applied to the subjects to return to subsequent meetings. The number of meetings actually attended was assumed to reflect their attraction to the group.

Starting the Manipulation of a Variable

It is possible and sometimes necessary, to start the manipulation of an experimental variable at the time the subjects are recruited for the experiment. This can be done by providing various expectations for the subjects which will affect the attitudes with which they come to the experimental situation, or by collecting information which will later be used to manipulate a certain desired variable. We shall give some examples of the experimental manipulation of a variable which begins at the time of recruitment.

Several experiments (3, 14, 28) have varied attraction to the group experimentally by manipulating the degree to which the subjects expected they would like, and be congenial with, the other members of the group. At the time of recruiting, those who volunteered to be subjects were asked to answer a number of questions which concerned characteristics of themselves, characteristics which they liked in other people, and characteristics which they disliked in other people. No attention was actually paid to these data in setting up a group, but, because the subjects had provided such information, the experimenters were plausibly able to tell some groups that the members would like one another and be congenial and to tell others that they would not be very congenial. The results of such experiments showed that the manipulations were successful.

Schachter (26) in his experiment on the rejection of deviates wanted to manipulate attraction to the group on the basis of interest in the activity in which the group was to engage. When the subjects were asked to join one of the clubs, each club was described in detail. Those who desired to join filled out an information sheet on which they were asked to give ratings of how interested they were in joining each of the available clubs. Some groups were composed of subjects who were highly interested in joining that specific club (high attraction to the group), whereas other groups were composed of persons who had indicated relatively low interest in

'real' for the subject and a situation which is 'experimental' for him. All of the situations are, in a sense, 'real' for the subject, and all of them likewise, are experiments from the point of view of the investigator. Some examples from other fields of investigation may illustrate our point more clearly. If a psychologist does an experiment in discrimination learning using rats as subjects, the situation is obviously an experimental one for the investigator. For the rat, however, it is undoubtedly a very real situation. The maze or discrimination box is a place where he works and gets fed. The basis of the 'reality' of the experimental situation for the subject is somewhat less clear when humans are used as subjects. Thus, for example, in an experiment on level of aspiration the subject may come to the laboratory knowing he is to help in an experiment. He is given a series of tasks to perform and is asked, before each task, what he is going to try to score on the subsequent task. One may well ask, In what sense is this situation a real one for the subject? Certainly it is not 'real' in the sense that it is a situation similar to those which the subject encounters in the ordinary course of events, on the other hand, it is certainly 'real' in the sense that powerful motives are brought into play and strong forces are set up which act on the subject and determine his behavior in lawful ways. Thus the situation in which one places the subject can be 'real' for him in that it brings into play powerful forces regardless of whether or not it is cognitively an experimental situation for him.

If the situation is cognitively a real one for the subject it is probably easier to bring powerful forces into play. It may be more difficult to produce equally strong forces if the situation is cognitively experimental. In the latter case the strength of the forces which can be brought into play depends largely upon the relations between the subject and the experimenter, the motivations which made the subject decide to volunteer for the experiment, and his desire to cooperate. These forces can, in the proper circumstances, be quite strong. It is much easier to create a laboratory situation which is cognitively experimental for the subjects. To create a cognitively 'real' situation and still be able to control and manipulate variables successfully may require a great deal of subterfuge and much attention to technical details. If the subject sees through the subterfuge, the whole experiment may be invalidated.

each subject was written a letter stressing the importance of her coming to the experiment. On the day before the meeting, each subject was spoken to by telephone to make sure that she would be present. In spite of these efforts, only five subjects appeared in a number of groups. In most of these instances the subjects who had arrived agreed to wait while others who had volunteered were telephoned until an appropriate person was reached who agreed to come down immediately. By this procedure very few groups had to be discarded. In Pepitone's (25) experiment on group productivity, the situation was designed so that it was essential to have three subjects present in each group. The group was to work on a task which was divided into three parts, each of which had to be performed by one subject. The experimenter scheduled four subjects for each group. Occasionally only two subjects appeared and the group had to be canceled. Most frequently three subjects appeared. When all four came, the last one was taken aside, the situation was explained to him, and he was allowed to observe the experiment in progress.

THE CONTENT AND FORM OF THE EXPERIMENTAL SITUATION

The investigator must make a number of decisions concerning how the situation is to be structured cognitively for the subjects, in what kinds of activities they will engage, and with what attitudes they come to the experiment.

"Real or Experimental" Situations

The experimental situation can vary from one which is frankly experimental to a situation which, for the subjects, is a 'real' one. The pros and cons for the various possibilities within this range are by no means all clear. Good evidence is lacking concerning which types of experimental situations are superior for which purposes. We shall, however, discuss some of the considerations which might lead an experimenter to set up his groups in one or another manner.

To discuss these advantages and disadvantages we must explain somewhat farther the distinction between a situation which is

real aspects of the situation, it was also not possible for the experimenter to engage in any further manipulation of variables while the meeting of the group was in progress. These problems, in this experiment, were satisfactorily solved by the use of paid participants, a technique which will be described later.

In an experiment on the effects of knowledge of religious affiliation Festinger (10) decided to use a cognitively experimental situation. This decision was made because it was obviously of importance to control the group session firmly and to carry on manipulations of variables while the session was in progress. The group consequently met with the knowledge that it was helping in an experiment. They were told to imagine that they were a club. There is no doubt that the forces in this situation were weaker than the forces which would have operated had the subjects actually been members of a club engaged in the same procedure. By virtue of the cognitively experimental aspects of the situation, however, this disadvantage of weaker motivation was counterbalanced by the precision of measurement and the control of extraneous variables.

The Choice of Activity for the Group

The choice of the activity in which the group once assembled in the laboratory, is to engage is somewhat dependent upon the decision concerning the cognitive reality of the experiment. There is, of course, much leeway in the choice of activity, although it must be one which is consistent with the purposes of the experiments and does not conflict with the other experimental decisions which have been made. If the experimental situation is to be cognitively real, there are restrictions on the type of activity which can be employed. If the situation is to be cognitively experimental, there is much less limitation and the selection of an activity which is well suited to the experimental purposes is easier. The activity must be chosen to allow for the manipulation of the variables, the collection of the measures in which the investigator is interested and the arousal of sufficiently strong forces so that the effects will be measurable. It is impossible, of course, to list all of the various activities in which laboratory groups may engage. We shall present a few examples of different kinds of activities which have been used and the reasons for their use.

We have then these relative advantages and disadvantages which the experimenter must consider when deciding whether to make the experimental situation cognitively experimental or cognitively real for the subject. If the experiment is cognitively real it will be easier to make it motivationally strong. On the other hand if the situation is cognitively experimental it will be easier to set up with an adequate amount of control and precision. The examples below illustrate the kinds of decisions which have been made on this question.

Lippitt (23) in his experiment on the effects of the behavior of autocratic and democratic leaders chose to make their experimental situation cognitively real for the subjects. To do this he organized school age children into clubs which had their club rooms in the investigator's laboratory. The experimenter functioned as the adult leader of these clubs. In this role he was able to manipulate the desired variables. Because of the desire to maintain a cognitively real situation the possible variations in the leaders' behavior were also limited. The differences between conditions that were produced were rather gross. It is possible that the lack of control and precision in this experiment offset the advantages gained by having a cognitively real situation.

Schachter (26) in his experiment on rejection of deviates also chose to have a cognitively real situation for the subjects because the major measures of rejection were to be obtained from verbal responses to questions. The investigator felt that these responses would have more validity if they were commitments to action on the part of the subject rather than answers to hypothetical questions. To obtain a cognitively real situation he organized clubs of college students.

Once more a major difficulty was the restriction on the manipulation of variables. Manipulations had to be devised which were consistent with the notion of a bona fide club. To create groups with high and low cohesiveness the investigator first ascertained the degree of interest of the subjects in each of two kinds of clubs and then manipulated the attraction to the group by composing some groups of persons who were all highly interested in the activity and other groups of persons who were only mildly interested. This type of manipulation of a variable by selection is probably not so satisfactory as other techniques would be. Because of the cognitively

experiment which the investigator can communicate to the subjects and which they will accept. If this is not done, the subjects usually conjecture about it and make guesses as to the true purpose. If a plausible orientation is not given, this important aspect remains uncontrolled.

The orientation which the experimenter gives the subject at the beginning should be plausible and should remain plausible as the experiment progresses. It is usually important that this plausible orientation *not* reveal to the subject the true focus of the experiment. The true purpose of the experiment and the true focus of the investigator's interest can, and should, be revealed to the subjects at the conclusion of the experiment.

TECHNIQUES FOR THE CONTROL AND MANIPULATION OF VARIABLES

Since the basic purpose of a laboratory experiment is to achieve a simple situation in which certain variables can be well controlled while others can be varied at will, we shall attempt, in the present section, to be as detailed as possible. We shall illustrate not only the various techniques which have been developed for controlling and manipulating variables but also the kinds of variables which have been successfully controlled and manipulated in the laboratory.

Use of Pre-experimental Instructions

The most obvious technique for controlling or manipulating variables is the use of pre-experimental instructions to the subject. Such pre-experimental instructions vary greatly in their effectiveness. It is probably safe to say that instructions to the subjects will be successful in manipulating variables when these instructions are kept simple, are given emphatically, and are plausible in the sense of being integrally related to the experimental activity in which the subjects are to engage. The major dangers in the use of instructions as a device for manipulating variables are (1) the possible inattention of the subjects when the instructions are given and (2) the possible variability from subject to subject in interpretation of the instructions. Because of these difficulties, it is probably unde-

Perhaps the most frequently employed group activity is discussion. Such an activity may be chosen when the purpose of the experiment is either to study the involvement of people in an activity, the amount of participation in an activity, or the communication or influence process that goes on in groups, or to provide a relatively interesting activity which will involve the subjects in order for the experiment to accomplish some other purpose in the meantime. Any topic which will be interesting to the subjects is suitable. The discussion may concern differences in opinion, as in the experiment by Back (3), it may be directed toward solving a problem, as in the experiment by Deutsch (9), or it may involve a sharing of experiences, as in the study of Festinger, Pepitone, and Newcomb (15).

When children are used as subjects, a play activity may frequently be appropriate. Thus, Thibaut (30), when he endeavored to create privileged and underprivileged subgroups, had one subgroup play an interesting and enjoyable game while the other subgroup took the role of helpers and servants to those who were actively engaged in having fun. Lippitt (23), in his experiment on autocratic and democratic leader behavior, used various games and craft activities which were appealing to school age children.

It is also possible to use work situations as the activity for the group. Kelley (19) felt that a work situation would be more conducive to the establishment of a status hierarchy, so in order to create a two level status hierarchy he used a work task in which the subjects had to arrange bricks according to a certain pattern. Pepitone (25), in an experiment on group productivity, used a work task which was constructed so that measures of production would be relatively easy to obtain.

These are but a few of the many possible examples of activities that can be prepared for a group. There is almost limitless room for the experimenter's ingenuity to create a situation which will be best for his experimental purpose.

The Orientation of the Subjects

Related to both the cognitive nature of the situation and the activity in which the group is to engage is the problem of what orientation to give the subjects in the experiment. It is highly desirable to have some plausible and understandable purpose for the

the personal goals that could be achieved through membership. This was done by informing the subjects that there was (or was not) a reward that would be given as a prize to each of the members in the best group.

These instructions were probably moderately successful. On the one hand, they were not integral to the experimental task. That is, the subjects could have done everything the experimenter required of them without these instructions ever having been given. The possibility of winning a reward or the likelihood that members would get along well with others in the group was, however, relevant to fairly important motives in the subjects. They probably were concerned about whether or not they would like the other persons and be liked by them. The possibility of a reward probably added to the motivation to do well in the eyes of the experimenter. The results of the experiment show that a difference between high and low attraction was created by means of these instructions.

In an experiment on the direction of communication in a group, Festinger and Thibaut (12) wanted to manipulate the subject's perception of the homogeneity or heterogeneity of the group. To create the perception of homogeneity, groups were told that the members had been carefully selected so that they were all in the same year in college and had equal interest in, and knowledge about, the problem they were to discuss. To create the perception of a heterogeneous group, they were told that great differences existed among them in their knowledge about and interest in, the problem under discussion. The manipulation of the variable by these instructions was only mildly successful. Probably few of the subjects were much concerned with whether the group was homogeneous or heterogeneous. Although differences between these conditions were obtained in the results, these differences were by no means strong. It might be expected that a more adequate manipulation of these variables would have produced much larger differences between the conditions.

In an experiment by Festinger *et al* (14), an attempt was made to manipulate three variables simultaneously, all by means of verbal instructions at the beginning of the experiment. The investigators were interested in the interaction among the variables of attraction to the group, perception of whether or not there were experts in the group, and perception of whether or not there was a correct

sirable to manipulate more than one variable at a time through the use of pre experimental instructions. Instructions which attempt to manipulate several variables simultaneously are likely to become so complex and so long that they render the manipulation ineffective. We shall illustrate the problems involved in the use of instructions by giving examples of successful and unsuccessful attempts at manipulating variables in this manner.

Deutsch (9) in his experiment on competitive and cooperative groups produced competitive or cooperative situations by differential instructions to the groups. In the competitive groups he told the subjects that all the members would be ranked according to their contributions in solving the problems given to the group and that their grades in the course would depend in part upon these rankings. It was explained that, thus, the one in that group who contributed most, irrespective of how the group as a whole performed, would get the highest grade and the one who contributed least would get the lowest grade. In the cooperative groups the experimenter told the subjects that their group was going to be compared with other groups that everyone in the group would receive the same grade and that this grade would be determined by how well the group as a whole did. These instructions were successful in creating the required conditions and they provide a good example of how instructions can be integrated into the experiment. They were successful because they provided essential explanation of the situation to the subjects—they defined the goals for the subject and defined the manner in which these goals were to be reached.

Back (3) in his experiment comparing groups of high and low cohesiveness wanted to vary the attraction to the group by using several kinds of motivation. In some groups he wanted to create in the members high or low attraction on the basis of personal liking for the other group members. To create high attraction he told the subjects that on the basis of the information they had written down when they volunteered he had matched people in this group so that he was quite sure they would be congenial and like one another. To create low attraction he told subjects that because of time scheduling difficulties he had been unable to match them very well but that he did not think they would dislike each other. In other groups attraction to the group was made dependent upon

was made to manipulate simultaneously three variables by verbal instructions to the subjects, a fourth variable was manipulated successfully by means of false reporting to the subjects. The subjects were to have a discussion among themselves concerning an issue about which each of them had already formed an opinion. Before the discussion some subjects were given the impression that the group overwhelmingly *agreed* with their own opinion on the issue whereas other subjects were given the impression that the group overwhelmingly *disagreed* with them. This was done in the following manner. Each subject wrote, on a slip of paper, his opinion on the issue which was to be discussed. Subjects were told that the experimenter would tabulate these and then give each person a tally which would show the opinion of each person in the group. Thus, knowing everyone's opinion, they would be able to proceed sensibly with their discussion. The tally which was handed to each of the subjects was entirely fictitious. Each of the subjects in whom the perception of group agreement was to be created was given a tally which showed all but one of the subjects agreeing very closely with him. Each of those in whom the perception of disagreement with the group was to be created was handed a tally sheet which showed everyone in the group at least two opinion steps removed from his own opinion. This false reporting proved successful in varying the degree of perceived agreement with the group.

We shall conclude the discussion of the technique of false reporting to subjects with an illustration of an unsuccessful attempt. Festinger and Hymovitch (13) attempted to create in subjects a feeling of rejection by the group. Four subjects, strangers to one another, met in the laboratory and were told that they were to work on a task which required cooperative effort although the various parts of the task would be divided among them. They were first to have a brief discussion among themselves and get to know one another so that they could decide how they wanted to organize the task. They were told that people who liked one another worked more productively together. Consequently, if there was any one in the group that they disliked, it would be better to exclude that person from the group. After the discussion, the subjects were given ballots on which each could indicate whether he wanted to work together with all the others or wanted to eliminate a member from the group. If subjects chose the latter alternative, they wrote down

answer to the discussion problem. This attempt to manipulate all three variables by pre-experimental instruction was not very successful. The amount of instruction which had to be given to the subject and the complexity of the instructions rendered them rather ineffective. It probably would have been better to manipulate one of these variables by instructions and to have devised techniques for manipulating the other two in other ways. We shall discuss below such other techniques of manipulating and controlling variables.

Use of False Reporting

False reporting to the subjects of the results of votes or of sociometric choices and the like is another technique for control and manipulation of variables. Such false reporting must always be done in a manner which will make the report appear plausible. If sufficient care is used to ensure the acceptance of the report as true, this can be an effective means of manipulating some kinds of variables.

Festinger (10) in his experiment on the voting behavior of Catholics and Jews in mixed groups used the technique of false reporting to the subjects to keep the situation identical for all groups. The members of the group voted for officers of the club in the following manner. There was first a nomination ballot to select two candidates for the election. The members of the group who received the most votes were to be the candidates in the final election. This nomination ballot was tabulated by the experimenter and, since the ballots were secret, it was simple for him to report falsely which two members had won the nominations. In this manner the experimenter was able to control which two persons were the candidates in each election. This experiment also employed paid participants (the use of which will be elaborated below) who were members of every group. By means of the false reporting of the results of the nomination ballot, the two candidates for each election in every group were two of the paid participants. One of the two candidates in each election identified herself as Jewish and the other identified herself as Catholic. Each election in each group was, thus, a standard situation.

In the experiment by Festinger *et al* (14) in which an attempt

Pepitone (24) reports an experiment in which he investigated the determinants of the perception of authority and approval in people. He was faced with the problem of how to provide a standard social situation for his subjects in which it would be meaningful to ask them for their perceptions of authority and approval. Using school-age children as his subjects, he let it be known in the school that, as part of a survey on interest in athletics, a three-man board would arrive in a few days to interview many of the students. Those who successfully answered the questions asked by the three-man board would win tickets to a college basketball game. The three-man board which came to the school and interviewed students individually consisted actually of three paid participants who had been trained by the experimenter. Scripts for each of the three had been carefully written so that each boy who was interviewed was asked exactly the same questions. The responses to the boys' answers were also standard for each of the conditions. In different conditions, however, the experimenter created authority differentials among the three board members and also differences among them in the extent to which they openly voiced approval of the boy who was being interviewed. The boy's perception of the relative authority and approval among the board members could be ascertained in an interview with each boy directly after his appearance before the three-man board. Thus, the experimental situation was effectively standardized.

Schachter (26), in his study of rejection of deviates, had three paid participants in each group. The topic for discussion was chosen so that all of the subjects would have opinions which very nearly agreed with one another. Paid participants were used to create various conditions of deviation from this group norm. One paid participant voiced an extremely deviant opinion and held to it throughout the discussion. Another paid participant voiced a deviate opinion at the outset but allowed himself to be influenced so that, in the end, he agreed with the other subjects. The third paid participant agreed at the beginning and continued to agree with the modal opinion in the group. Thus, standard conditions of deviation from the group norm were achieved and, by rotating the paid participants among the various roles from group to group, it was also possible to equate for personality factors. We must emphasize that these paid participants had been very carefully trained in how

the name of the member they wanted to reject. Each subject was then taken to a separate room and was told that the experimenter would tell him the results of the ballot as soon as possible. Each subject was then privately told that the others had unanimously voted to reject him.

This false report to the subject was rarely successful. The overwhelming majority of the subjects refused to accept it and immediately suspected that the experimenter was not telling the truth. The reason for the failure were probably twofold. The experience with the others in the preliminary discussion did not provide grounds on the basis of which they could accept the reported rejection. Also the false report was unpleasant enough so that the subjects did not want to accept it. Many subjects refused to accept the report even though they could not verbalize any reason for suspicion or disbelief. This technique had to be abandoned in this experiment.

Use of Paid Participants

The use of paid participants who are part of the experimental group and are accepted as such by the subjects is a powerful technique for the control and manipulation of variables. It is, however, a relatively expensive and tedious procedure. When paid participants are used, the details of their behavior must be exactly planned in advance and much time must be spent training and rehearsing them. We shall give some examples to illustrate the great variety of uses to which such paid participants may be put.

A relatively simple and effective use of paid participants to manipulate a variable is found in an experiment by Sherif (29). The same technique has been used by others for the same specific purpose (6). These experiments brought two persons together in the laboratory so that the degree to which the judgments of one would influence the judgments of the other might be investigated. The subjects were asked to judge the amount of movement of a point of light. This autokinetic effect (the light does not actually move) provided a rather ambiguous stimulus. These experiments used as one of the group members a paid participant who, by making a standard prearranged series of judgments, was able to produce a standard situation for all subjects with specified differences between his judgments and the subjects' initial judgments.

before it was delivered to another group member, so that the whole communication process could be reconstructed in the analysis. Aside from these measurement problems, there were other reasons for restricting the discussion to written notes. In an oral discussion, the person who is talking may be primarily addressing one or two others in the group, but, whether he likes it or not, what he is saying is simultaneously heard by everyone. This introduces additional complexities. By limiting the communication process to written notes, with the further restriction that each note could be sent to only one person, the situation was kept simple and manageable. A further difficulty in using an oral discussion for the purposes of this experiment is the marked tendency for people to answer when remarks are addressed to them. This is fully demonstrated by the usually high correlation obtained between the number of times a person communicates to others and the number of times he is the recipient of communication (17). Since the experimenters were concerned primarily with other determinants of the direction of communication, this would have been a complicating factor. The further restriction that the written notes could not be signed avoided this complication. The recipient of a note did not know from whom it came. The pads of paper on which the subjects wrote their notes were marked so that later, in the analysis, the experimenter could tell who had written each note as well as to whom each note was addressed.

In his experiment on communication in a status hierarchy, Kelley (19) also restricted communication to written notes. Again there were a number of functions served by this restriction on the communication process. First, the experimenter intercepted all the notes written and thus had a detailed record of the communication process. Secondly, since all communication was by written notes, the experimenter could easily manipulate the communication process. Actually, none of the notes which the subjects wrote to one another was delivered. The notes which they received were fictitious ones designed to produce certain effects. In this manner a standard pattern of receiving communications from others was established for every group in all of the experimenter's conditions.

Restrictions on the behavior of the group can also be produced by an appropriate activity in which the group must engage. An activity can be chosen to eliminate certain complications, restrict

to behave in the group and in what kinds of things they could and could not say

In the study by Festinger (10) of the effect of knowledge of religious affiliation four paid participants were members of every group which met. These paid participants were relied upon to control many variables and to create a standard situation. In the middle of the experiment, when everyone was identified according to her name and religious affiliation two of these paid participants announced that they were Catholic and two announced that they were Jewish. The ones who said they were Jewish or Catholic were rotated from group to group so that actual religious affiliation and personality differences were equated among all the conditions. In this manner, many powerful variables, which would affect preferences for people, were controlled and the effects of knowledge of religious affiliation were permitted to emerge quite clearly.

The three foregoing examples of the use of paid participants in laboratory experiments hardly demonstrate adequately the possible range of uses to which this technique may be put. With sufficient ingenuity on the part of the experimenter and sufficient time in planning the behavior of the paid participants and in adequately training and rehearsing them, very powerful effects can be produced. There is ample evidence of the success of the control and manipulation of variables with the aid of paid participants.

Restriction of Behavior Possibilities

It is possible to exercise control over a situation and to manipulate variables by creating a situation which restricts the possibilities of behavior.

Festinger and Thibaut (12), in their experiment on the determinants of direction of communication, restricted the group to the use of written notes in carrying on their discussion. This decision was made for a number of reasons. If the discussion had been an oral one, the direction of communication (who spoke to whom) would have had to be recorded by observation of the group while the discussion was in progress. Such observation in fairly large groups is difficult and sometimes quite unreliable (see Chap. 9). By the use of written notes, a permanent record was immediately available. The exact time each note was written was recorded on it.

exhaustively the various kinds of techniques for the control and the manipulation of variables. Those described are no more than a few examples of the wide variety of which an experimenter can avail himself. Many more possible techniques are likely to be developed in the near future. It should again be stressed that when one employs new techniques for manipulation of variables, or even some of those already developed, it is important to conduct preliminary experimentation to make sure that the manipulation is actually working.

OPPORTUNITIES FOR MEASUREMENT IN LABORATORY EXPERIMENTATION

Opportunities for collecting data in a laboratory experiment are present at all phases, from the recruiting of subjects until the end of the experimental sessions. There are, of course, some restrictions on what kinds of measurement can be employed at various phases in this process. These depend upon the design of the experiment and the way in which it is cognitively structured for the subjects. We shall point out some of the measurement possibilities at each of the stages of a laboratory experiment.

The first opportunity for measurement occurs before the experimental session takes place. Such measurement may be made at the time of recruiting subjects or when the subjects have assembled in the laboratory but before the experiment has begun. The exact time at which the measurement is done is immaterial and is generally selected for convenience. Such measurements, using a questionnaire or an interview, can have the following purposes: (1) to obtain some measure which will be compared to a similar one taken during or after the experiment; and (2) to enable the experimenter to control a variable by manipulating the composition of the group according to these measures.

In some experiments, it is essential for data to be collected before the experiment began. Thibaut (30), in his experiment on the cohesiveness of privileged and underprivileged subgroups, employed pre-experimental measurements to equate groups in the experiment and also to have a comparison between a pre-experimental and a postexperimental measure. The subjects were members

the range of behavior, or produce certain reactions in the subject French (16) in his experiment on the effects of frustration and fear on organized and unorganized groups, produced frustration in his groups by means of the activity in which they engaged. The groups were put to work on a task which was impossible to complete. The frustration engendered in this manner was unmistakable.

In his experiment on the relationship between influence and group cohesiveness, Back (3) wanted to produce a situation in which two subjects meeting together, had different interpretations of, or opinions about, the same set of facts. Before they came together, each subject was given a set of three pictures and asked to write a story about them. Each of the subjects was actually given different pictures, which would force different interpretations. The differences between the sets of pictures, however, were so slight that none of the subjects ever suspected that he had seen different pictures. In this manner, by appropriate choice of activity, Back was able to ensure that, in every group, there would be a difference of opinion between the two subjects at the beginning of their discussion.

In experiments by Bavelas (5) and his colleagues (20) on the effectiveness of different patterns of communication in groups, a technique has been employed which is perhaps the most extreme example of restriction in a situation. In these studies the experimenters were concerned with determining which of a number of patterns of communication among members of a group would result in more effective problem solving. To produce the different patterns of communication, the experimenters allowed some members to communicate to one another and prevented others from doing so. By this simple restriction on which channels of communication were or were not available various communication patterns were established. In these experiments the purposes of the investigators and the artificiality of the manipulation device were not hidden from the subjects. The restriction of the situation, however, was such that the subjects had to behave within it as well as they could. The results of these experiments show that the manipulation was successful. Such extreme and frank restriction of the situation would be appropriate, of course, only for a relatively selected range of problems.

In the foregoing discussion, we have by no means covered

Questionnaires and interviews may also be used during the course of the experiment. These may take the artificial form of questions interpolated into, and momentarily interrupting, the experiment or they may be disguised as election votes or expressions of opinion necessary to the conduct of the experiment.

Schachter (26), in his experiment on rejection of deviates, created a situation which was cognitively real to the subject. The groups were clubs which the subject had joined and which the subject expected would continue meeting periodically. It was fitting, consequently, to ask the subjects to elect committees to carry on various of the club functions and to vote on when and how often the club should meet. In this experiment, the data collection was seen by the subjects not as such but rather as part of their functioning as members of a club.

In the Festinger (10) experiment on mixed Catholic and Jewish groups, the major data were collected by holding elections for officers of a club. Here the situation was cognitively experimental for the subjects and the voting was undoubtedly seen as part of the experimental procedure. The results indicate it to have been an adequate method of data collection.

One can also collect a wide variety of data by questionnaires, interviews, or tests at the conclusion of the experimental session. The techniques of such data collection are discussed in Chapters 8 and 9.

SUMMARY

Laboratory experiments constitute a powerful technique for investigating relationships among variables. The essence of such experiments may be described as observing the effect on a dependent variable of the manipulation of an independent variable under controlled conditions. Such experiments, if well designed, can produce clear and unambiguous results which may add to a theoretical body of knowledge.

It is important to remember, however, that laboratory experimentation, as a technique for the development of an empirical body of knowledge, cannot exist by itself. Experiments in the laboratory must derive their direction from studies of real-life situations,

of already existing clubs. The investigator met the group at some designated place, usually their Y M C A or their club. He provided transportation for them to the experimental rooms. Before setting out for the laboratory, he asked them to answer a questionnaire concerning who their friends were among the other boys. He then brought them to the experimental rooms and was able to divide them into two subgroups so that each person had about as many of his friends within his own subgroup as in the other subgroup. After the experiment was concluded, the boys were again asked to answer the same sociometric questions. In this manner the investigator was able not only to equate his subgroups for amount of friendship within them but also to provide a basis for determining the effect of the experimental procedure on this variable.

Most of the possibilities for measurement occur, of course during the actual progress of the experiment. One of the most frequently used measurement devices is observation of the group as it carries on its activities (dealt with in detail in Chap. 9). We shall discuss here some of the other kinds of data collection which are possible during the experiment.

The product of the activity in which the group engages is a major source of data. This product may take any of a variety of forms and may be analyzed in various ways by the investigator.

Kelley (19) in his experiment on communication in a status hierarchy, had his subjects arrange bricks in a certain pattern on the floor in accordance with instructions communicated to them. The actual product—that is, the exact pattern of bricks with which the group finished—was recorded by the experimenter and was used to obtain a measure of adequacy of production.

In his experiment on competitive and cooperative groups Deutsch (9) had the subjects discuss, and write solutions to various human relations problems. He then analyzed these written products of the group discussion to obtain measures of the adequacy of the solution to the problem.

Closely related to such products are various records which the subject makes in the process of doing the required activity. Thus in the Kelley (19) experiment and in the Festinger and Thibaut (12) experiment on direction of communication, the actual notes which the subjects wrote while carrying on the discussion were the main source of data.

- 5 Bavelas, A Communication patterns in task-oriented groups In Lerner, D, and Lasswell, H D (eds) *The policy sciences* Stanford Stanford Univ Press, 1951, pp 193 202
- 6 Bray, D W The prediction of behavior from two attitude scales *J Abnorm Soc Psychol*, 1950, 45, 64 84
- 7 Coch, L, and French, J R P. Jr Overcoming resistance to change *Hum Relat*, 1948, 1, 512 532
- 8 Dashiell, J F An experimental analysis of some group effects *J Abnorm Soc Psychol*, 1930, 25, 190 199
- 9 Deutsch, M An experimental study of the effects of cooperation and competition upon group process *Hum Relat*, 1949 2, 199 232
- 10 Festinger, L. The role of group belongingness in a voting situation *Hum Relat*, 1947, 1, 154 181
- 11 ———, Cartwright, D, Barber, K, Fleischl J, Gottsdanker J, Keaven A, and Leavitt G A study of rumor its origin and spread *Hum Relat*, 1948, 1, 464-486
- 12 ———, and Thibaut, J Interpersonal communication in small groups *J Abnorm Soc Psychol*, 1951, 46, 92 99
- 13 ———, and Hymovitch, B *Communication as a consummatory activity* Unpublished manuscript
- 14 ———, Gerard, H B, et al *The influence process in the presence of extreme deviates* In press
- 15 ———, Pepitone, A, and Newcomb T Some consequences of deindividuation in a group *J Abnorm Soc Psychol*, 1952, 47, 382 389
- 16 French, J R. P, Jr Organized and unorganized groups under fear and frustration Univ of Iowa Studies Studies in Child Welfare Volume 20 *Studies in Topological and Vector Psychology III*, 1944 pp 229 308
- 17 ———, and Bradford, L (eds) *The dynamics of the discussion group* *J Soc Issues*, 1948, 4, 65
- 18 Jenness, A Social influences in the change of opinion the role of discussion in changing opinion regarding a matter of fact *J Abnorm Soc Psychol*, 1932, 27, 29 34, 279 296
- 19 Kelley, H H Communication in experimentally created hierarchies *Hum Relat*, 1951, 4, 39 56
- 20 Leavitt, H J Some effects of certain communication patterns on group performance *J Abnorm Soc Psychol*, 1951, 46, 38 50
- 21 Lewin, K, Lippitt, R, and White R Patterns of aggressive behavior

and results must continually be checked by studies of real life situations. The laboratory experiment is a technique for basic and theoretical research and is not the goal of an empirical science.

We have in this chapter, enumerated in some detail many techniques for designing laboratory experiments and for manipulating different kinds of variables in a variety of ways. Many of these techniques for the manipulation of variables involve deception, prevarication, misdirection of subject, and the like. As long as an investigator works with human subjects, it is impossible to overemphasize the necessity for keeping in mind the responsibilities to the subject and the ethics which the experimenter must follow. It is important if such experimentation is to continue and is to be tolerated by the people who help in it, that the experimenter perform a service to the subjects in exchange for their help. In all laboratory experiments it should be a firm policy to give the subjects a full explanation at the conclusion of each experiment. This sometimes requires spending more time explaining and discussing matters with the group than it took to do the experiment. If it is done well, the subjects leave feeling that they have learned something and have not wasted their time. The subjects do not resent having been misdirected and deceived if they can see the reasons for the deceptions and understand the purposes.

BIBLIOGRAPHY

- 1 Allport F H The influence of the group upon association and thought *J Exper Psychol* 1920 3, 159 182
- 2 Asch S E Effects of group pressure upon the modification and distortion of judgments. In Guetzkow H (ed) *Groups leadership and men* Pittsburgh Carnegie Press 1951 pp 177 190
- 3 Back K Influence through social communication *J Abnorm Soc Psychol*, 1951 46 9 23
- 4 ——— Festinger L Hymovitch B Kelley H Schachter S and Thibaut J The methodology of studying rumor transmission *Hum Relat* 1950 3, 307 312

- 5 Bavelas A Communication patterns in task oriented groups In Lerner, D, and Lasswell, H D (eds) *The policy sciences* Stanford Stanford Univ Press, 1951, pp 193 202
- 6 Bray D W. The prediction of behavior from two attitude scales *J Abnorm Soc Psychol*, 1950, 45, 64 84
- 7 Coch L, and French, J R P. Jr Overcoming resistance to change *Hum Relat*, 1948, 1, 512 532
- 8 Dashiell, J F An experimental analysis of some group effects *J Abnorm Soc Psychol*, 1930, 25, 190 199
- 9 Deutsch, M An experimental study of the effects of cooperation and competition upon group process *Hum Relat*, 1949 2 199 232
- 10 Festinger, L The role of group belongingness in a voting situation *Hum Relat*, 1947, 1, 154 181
- 11 ———, Cartwright, D, Barber, K, Fleischl J Gottsdanker J Keyesen A and Leavitt G A study of rumor its origin and spread *Hum Relat*, 1948, 1, 464 486
- 12 ———, and Thibaut, J Interpersonal communication in small groups *J Abnorm Soc Psychol*, 1951, 46, 92 99
- 13 ———, and Hymovitch, B *Communication as a consummatory activity* Unpublished manuscript
- 14 ———, Gerard H B, et al *The influence process in the presence of extreme deviates* In press
- 15 ——— Pepitone, A, and Newcomb T Some consequences of de individuation in a group *J Abnorm Soc Psychol*, 1952 47, 382 389
- 16 French J R P, Jr Organized and unorganized groups under fear and frustration Univ of Iowa Studies Studies in Child Welfare, Volume 20 *Studies in Topological and Vector Psychology III*, 1941 pp 229 308
- 17 ———, and Bradford, L (eds) *The dynamics of the discussion group* *J Soc Issues*, 1948, 4, 65
- 18 Jenness, A Social influences in the change of opinion the role of discussion in changing opinion regarding a matter of fact *J Abnorm Soc Psychol*, 1932, 27, 29 34 279 296
- 19 Kelley, H H Communication in experimentally created hierarchies *Hum Relat*, 1951, 4, 39 56
- 20 Leavitt, H J Some effects of certain communication patterns on group performance *J Abnorm Soc Psychol*, 1951, 46, 38 50
- 21 Lewin K Lippitt R and White R Patterns of aggressive behavior

and results must continually be checked by studies of real life situations. The laboratory experiment is a technique for basic and theoretical research and is not the goal of an empirical science.

We have in this chapter enumerated in some detail many techniques for designing laboratory experiments and for manipulating different kinds of variables in a variety of ways. Many of these techniques for the manipulation of variables involve deception, prevarication, misdirection of subject and the like. As long as an investigator works with human subjects it is impossible to overemphasize the necessity for keeping in mind the responsibilities to the subject and the ethics which the experimenter must follow. It is important if such experimentation is to continue and is to be tolerated by the people who help in it that the experimenter perform a service to the subjects in exchange for their help. In all laboratory experiments it should be a firm policy to give the subjects a full explanation at the conclusion of each experiment. This sometimes requires spending more time explaining and discussing matters with the group than it took to do the experiment. If it is done well the subjects leave feeling that they have learned something and have not wasted their time. The subjects do not resent having been misdirected and deceived if they can see the reasons for the deceptions and understand the purposes.

BIBLIOGRAPHY

- 1 Allport F H The influence of the group upon association and thought *J Exper Psychol* 1920 3 159 182
- 2 Asch S E Effects of group pressure upon the modification and distortion of judgments In Guetzkow H (ed) *Groups leadership and men* Pittsburgh Carnegie Press 1951 pp 177 190
- 3 Back K Influence through social communication *J Abnorm Soc Psychol* 1951 46 9 23
- 4 ——— Festinger L Hymovitch B Kelley H Schachter S and Thibaut J The methodology of studying rumor transmission *Hum Relat* 1950 3 307 312

PART II

Procedures for Sampling

Before we proceed from research settings to the collection of data, it is necessary to pause for a consideration of sampling—how it is done and what its implications are. The question of sampling may be simply stated: how is the investigator going to decide what persons or groups or organizations or communities to use for the collection of his data? The way this decision is made will affect the conclusions which may be drawn and the precision of these conclusions.

Many investigators may protest that they do not make these decisions—that these decisions are made for them. One does research in industries into which one can get entree, one uses as subjects in laboratory experiments those people who volunteer, and the like. But such situations, frequent though they may be, do not obviate

- in experimentally created social climates *J Soc Psychol* 1939 10 271 299
- 22 Libo L *The use of a projective device to measure attraction to a group* Ph D thesis Stanford University 1951
- 23 Lippitt R An experimental study of the effect of democratic and authoritarian group atmospheres *Univ of Iowa Studies Studies in Child Welfare Studies in Topological and Vector Psychology I* 1940 pp 45 195
- 24 Pepitone A Motivational effects in social perception *Hum Relat* 1950 3 57 76
- 25 Pepitone E *The productivity of groups* Ph D thesis Univ of Michigan 1952
- 26 Schachter S Deviation rejection and communication *J Abnorm Soc Psychol* 1951 46 190 207
- 27 ——— *Group derived restraints and audience persuasion* In press
- 28 ——— Ellerston N McBride D and Gregory D An experimental study of cohesiveness and productivity *Hum Relat* 1951 4 229 238
- 29 Sherif M An experimental approach to the study of attitudes *Sociometry* 1937 1 90 98
- 30 Thibaut J An experimental study of the cohesiveness of under privilege groups *Hum Relat* 1950 3 251 278

Selection of the Sample

Leslie Kish

There is no hard and fast rule for deciding just where the specialized work of "sampling" begins and ends, the term has been used with various meanings in different contexts. It seems convenient to exclude the processes and problems of making observations (or measurements) from the area of this chapter. But it must be emphasized that this exclusion is arbitrary—that the problems of errors of response and nonresponse have bearing on the sample design (see Section 16).

Assume, then, that a method of observation has been decided on whereby the value of some characteristic may be obtained for any of the N elements (members, individuals, cases) which comprise the population (or universe). A target of empirical research is usually some numerical expression which summarizes the information about the characteristic from all the N elements of the entire population—a parameter, a population value. An example of such a parameter is the mean $\bar{x} = \frac{1}{N} \sum x$ *. Moreover, the mean is an important and commonly known and used parameter. Therefore, and for convenience and brevity, much of the discussion to follow will center around it. However, many of the general remarks concerning the mean will apply also to other parameters.

To obtain the exact value of a parameter, observations have to be

*The number of elements in the population is denoted here by N . The symbol x stands for the values of the variable characteristic. Then there are N separate values denoted by the general term x . The symbol $\sum x$ stands for the sum of the N values of x —that is, for $x_1 + x_2 + \dots + x_N$.

the necessity for a consideration of the nature of the sample and its characteristics

Sampling theory has made enormous strides in recent years, mainly in connection with the problems arising from large scale survey operations. It is, consequently, easiest to talk about sampling in connection with the problems of surveys, and it is easiest to see the applications in that context. But the applications exist elsewhere, too, and they must be discovered and used. As a single example, the reader may notice that the discussion of cluster sampling in the following chapter is quite relevant to laboratory group experiments where perhaps twenty groups of six persons each are the subjects of data collection.

employees? We obtain a (payroll) list numbered from 1 to 12,000 on which each of the 12,000 employees appears once. From a table of random numbers, 400 different¹ numbers are drawn (4 p. 34). At each draw a five-digit number not greater than 12,000 is taken. These 400 numbers will designate the numbers of 400 employees. Their names and other necessary identification are obtained and given to interviewers. They are interviewed, and their answers are reduced to a numerical code.

Now we want to estimate the proportion of all the employees in the factory who hold a certain attitude, that is, we shall calculate a sample value to serve as an estimate of the population value. For simple random sampling, the sample mean is obtained by the simple and familiar procedure of adding up all the values of the sample cases and dividing by their number n . In symbols²

$$\bar{x}_o' = \frac{1}{n} \sum x$$

A proportion p is only a special kind of mean (\bar{x}) where all the Np elements which possess a specified attribute are denoted by the value of $x = 1$ and all other $N(1 - p)$ elements are assigned the value $x = 0$. The sample mean is obtained by dividing the number in the sample (r) possessing the attribute by the total number in the sample. $\bar{x}_o' = p' = r/n$. For example, let us say that in the 400 interviews there were 80 "yes" answers to a question. The sample proportion $p' = 80/400 = 20\%$ is our estimate of p , the proportion that would have been obtained had the entire population of 12,000 been interviewed by the same procedures. Moreover, our estimate of the aggregate, of the *total number* (Np) of employees who would have said "yes" to that question, is $Np' = 12,000 \times 20\% = 2,400$.

The sample mean obtained from a single sample is only one of many possible values that could have been obtained, it is subject to sampling variability. We want to estimate the population value \bar{x} . To be able to do this, we must have a measure of the variability to which

¹By the word *different* we mean that if one of the numbers comes up a second time we disregard it. Thus, at each choice we are selecting among the still unselected elements.

²The subscript o is used here to denote that the sample design was simple random sampling. Similarly, estimates derived from other designs will also be designated by specific subscripts.

made on all the elements in the entire population. However, seldom can information be obtained about all members of a population large enough to be interesting. Usually practical considerations, particularly of cost, force us to be satisfied with making inferences about population values from data which are based on a sample only. What are the tasks facing the researcher in planning a sample in order to make inferences about the population values? He must (1) select a sample, (2) make observations, and (3) compute the *statistics*—that is, estimates based on the sample data.

These sample values are of little interest in themselves. They are worth obtaining and are of interest to us only in so far as they yield information about the corresponding values in some population. Thus, the selection of the sample and the process of estimation are two tasks which are incurred jointly because the population values are estimated from a sample. Let us call the joint procedures of selection and estimation the *sample design*.

This chapter will concentrate on the less mathematical aspects of different procedures of selection. The important but somewhat technical problems of the use of different kinds of estimation will be largely neglected except for a brief discussion in Section 22. In the illustrative material the method of observation is assumed to be the interview. However, the procedures have general applicability to a great variety of other procedures, situations, and problems.

FUNDAMENTALS OF SAMPLING

1 *A Simple Random Sample*

Before proceeding further, let us look at an illustration of a specific sample design. Suppose that it is desired to learn something about the attitudes of the 12,000 employees of a factory. Suppose, also, that it has been decided that the method of sampling is to be "simple random sampling" and that the size of the sample will be 400 interviews.[†]

How do we select a "simple random sample" of 400 out of 12,000

[†]At this point, some may think of the widely used sampling procedure whereby every k th (here every 30th) element is selected into the sample. That distinct method of sampling, involving the use of intervals of selection, we call 'systematic sampling' and discuss in Section 10.

$$\text{est s e of } \bar{x}' = 983 \times \sqrt{\frac{20(80)}{399}} = 983 \times 0.200 = 0.197$$

The estimated standard error of the aggregate $N\bar{x}'$ can be given as

$$\text{est s e of } (N\bar{x}') = N (\text{est s e of } \bar{x}')$$

Hence, the standard error of the estimated 2400 employees who would have said "yes" to the question is $12,000 \times 0.197 = 236$ employees

The simple random sample as described here is seldom used in practice. It occupies a central place in sampling theory because it serves as a standard of comparison and because it is the basis for the various modifications of more complicated designs superimposed on it. In addition, its treatment here is justified by the fact that most of the formulas used in introductory texts, with which the reader is assumed to be acquainted, refer to samples obtained by simple random sampling. The reader is warned that those formulas can not be used validly in connection with samples obtained from other kinds of sample designs.

2 *The Sampling Distribution of the Estimate*

The information that the mean of a specific sample is 20 has no intrinsic practical worth, its value lies in that it tells us something about the population mean \bar{x} . Once we know that a sample mean \bar{x}' is 20, just what can we say about \bar{x} ? We know that the unknown \bar{x} will differ from our known $\bar{x}' = 20$, but we do not know by how much. The sample mean will depend on which sample of 400 persons happened to be selected, since different samples of 400 persons would give different sample means. The deviation from the population mean of any single sample mean is unknown, it may be large or small, plus or minus. The only fruitful way of looking at sampling variability is in terms of what size deviations are likely to occur in the long run. That is, we ask: given a sample design, what values of the sample mean \bar{x}'

*The discussion in this section concerns sample means derived from any kind of sample design. Here \bar{x} without any subscript denotes a sample mean in general, without specifying the sample design. The symbol \bar{x}' is used to denote any one of the values in the distribution of all sample means which may be obtained with the specific design (Section 16). The 'sample mean' here denotes the sample estimate of the population mean. It does not necessarily refer to the sample mean of the sample cases in Sections 8, 11, and 22; other formulas are given for the means of some other designs. The discussion is relevant for the sampling distribution of statistics other than the mean.

\bar{x}_o' is subject (Sections 2 and 3) The standard error, the square root of the variance, of the sample mean is the usual measure of that variability For simple random sampling, the variance of the sample mean can be estimated from the sample as

$$\text{est var of } \bar{x}_o' = (1 - f) \frac{s^2}{n}, \text{ where } s^2 = \frac{1}{n-1} \sum (x - \bar{x}_o')^2,$$

that is, the variance of the sample cases, with $(n - 1)$ used as divisor (The statistic s^2 is a sample estimate of the variance of the elements in the population)

This formula holds also, of course, in the event that our mean is a proportion However, in that event there is a form which saves the labor of squaring

$$\text{est var of } p' = (1 - f) \frac{p'(1 - p')}{n - 1}$$

It may be noted that here $\frac{n}{n-1} p'(1 - p')$ takes the place of its equivalent s^2

These formulas will appear familiar to the reader with the possible exception of the factor $(1 - f)$ which for a simple random sample is $(1 - f) = (1 - n/N)$ This "correction for a finite population" arises when sampling from a finite population "without replacement" This last phrase refers to the procedure which prevents any element from being selected into the sample more than once (We did this by selecting 400 *different* random numbers, not allowing the same number to appear twice) In most practical cases the sampling fraction (n/N) is so small that this correction is of no importance In reading the formulas, the reader should slip past it and on to the important parts of the formulas

In the present case the sampling fraction is $n/N = 400/12,000 = 1/30$ Then $(1 - f) = 1 - 1/30 = 29/30 = .967$ On the standard error, the effect is still less, being equal to $\sqrt{1 - 1/30}$ which is closely equal to $1 - 1/2 \cdot 1/30 = .983$ In most cases this factor is so close to unity that multiplying by it has no appreciable effect In our illustration we have for $p' = .20$

$$\text{est var of } \bar{x}_o' = .967 \frac{.20(.80)}{399} = .967 \times .000401 = .000388$$

Its square root is .0197 Or, again,

The deviation of each possible sample is taken from the mean of the sampling distribution \bar{x}'' . These deviations may be denoted as $(\bar{x}' - \bar{x}'')$. Then the variance of \bar{x}' around \bar{x}'' is calculated by taking the sum of the squared deviations each multiplied by its probability—its “relative frequency”—of occurrence

$$\text{var}(\bar{x}') = \sum P_i (\bar{x}'_i - \bar{x}'')^2$$

(The summation is for all possible samples)

The standard error of the sample mean \bar{x}' is defined as the square root of this variance. In other words, the sampling distribution represents the random fluctuation, the variability of the sample mean, due to a specified sample design. The amount of variability is measured in terms of the standard deviation of the sampling distribution. This very important quantity is called the standard error of the sample mean \bar{x}' , and we shall abbreviate it as *se* of \bar{x}' . Remember that it is always defined as the standard deviation of the distribution of all sample means under a specified sample design.

Now, as we have said above, it is not possible in practical work to obtain the standard error directly from the actual tabulation of the distribution. However, through the use of the concept of the sampling distribution, mathematical statisticians have developed specific useful formulas for the variances and standard errors of sample means for many practical sample designs. Moreover, they have developed practical formulas for obtaining *estimates* of the variances and standard errors of sample means—estimates which can be calculated from the data of a single sample. One example is the formula given for a simple random sample

$$\text{est se}(\bar{x}_o') = \sqrt{1-f} \frac{s}{\sqrt{n}}$$

The basic definition of the standard error of the mean is the same for any sample design: it is the standard error of the sampling distribution for that specified sample design. But it must be emphasized that for different sample designs there are different formulas of the estimated standard error, some of these will be given later in the sections devoted to different sample designs. They serve as powerful tools in the form of confidence intervals

are possible and what is the probability of occurrence of each of those values? The array of possible values of \bar{x}' , each with its probability of occurrence, is a distribution, it is the sampling distribution of the possible sample means \bar{x}' †

In the course of empirical research, when we calculate the mean from the single available sample, we cannot obtain the actual values for all the sample means possible under the design. Nevertheless, the distribution of all possible sample means is an important theoretical concept. It is what we should have in mind when we think about sampling fluctuations, sampling errors, standard errors, and such. Hence suppose that there has been specified a sampling design to be applied to a given population—including the size of the sample, the method of selection, and the method of estimation. Now imagine that all the sample means (\bar{x}') possible under the design are calculated, and so is the probability (P_i) of occurrence of each of those sample means †. This probability is analogous to the "relative frequency" in the calculations of the common formulas for variance. Now we have the sampling distribution of the different values that the sample mean \bar{x}' may take. The mean of that distribution—the mean of the possible sample means—is denoted by \bar{x}'' , and in well designed samples \bar{x}'' is either equal to the parameter \bar{x} or is close to it. The serious problems of the differences between \bar{x}'' and \bar{x} which arise in practice are discussed briefly in Section 16. Until then, we shall ignore the differences between the parameter \bar{x} and the mean of the sampling distribution \bar{x}'' ‡

†See (18 p 244 p 3743 and 16 p 95103)

‡Imagine that, after the sample design was specified, sample after sample was drawn. The mean for each sample was calculated and these values were tabulated. After many samples were drawn, the form of this distribution would gradually become more stable. As the number of samples drawn increased, the shape of the distribution would approach the true sampling distribution. In Section 3 we mention that for most large samples that distribution is close to a normal distribution.

§These differences exist, however, and are of great practical importance, owing to the presence of the nonsampling errors of response and nonresponse. Hence, in this chapter the statements of statistical inference are made not to the parameters but instead to population values. These are the values that would have been obtained if the entire population had been designated for observation rather than only a sample (Section 16). They are subject to the same sources of errors of response and nonresponse to which the sample estimates \bar{x}' are subject. They may be thought of as roughly equal to the mean \bar{x}'' of the distribution of the sample estimates.

(c) In the illustration in Section 1 we calculated the sample mean as .20 and its estimated standard error as .02. Now we make the statement that the population value lies in the interval between the value of $.20 - 2 (.02)$ and the value of $.20 + 2 (.02)$; that is, between 16 percent and 24 percent. This statement is the result of the sample we happened to draw; on the basis of another selection, we might have said between 19 and 25 percent, or between 14 and 24 percent, etc. The probability that statements of this kind are correct is 0.95. That is, in the long run, about 19 out of 20 statements of this kind will be correct. Thus, there are 5 chances in 100 that the statement would turn out to be incorrect; that, if we obtained the population value, we would find that it lies outside the limits we set (4, pp. 73-74).

In general, we make the statement that the population mean \bar{x} is somewhere between the value of $\bar{x}' \pm 2$ (est. s.e. of \bar{x}'); and that statement will have a 95 percent probability of being correct. (This assertion assumes that a *valid* estimate of the standard error was obtained, and that the approximation to normality of the sampling distribution is good enough.) We may choose our confidence intervals as $\bar{x}' \pm t$ (est. s.e. of \bar{x}'); the larger t is, the greater the probability that the statement is correct. Here t is the "normal deviate," which may be found in tables in many statistics textbooks.

Some values of t and the corresponding values of the probability of making correct statements, are:

t	.67	1.00	1.96	2.58	3.00	4.00
P	.50	.68	.95	.99	.997	.99994

The probability that we tell the truth increases rapidly with an increase in the value of t ; that is, with an increase in the length of the confidence interval. However, the longer the confidence interval, the less useful it is. It is general practice to fix the level of probability at some point and then use the corresponding t to get the length of the interval. In social science frequently the 95-percent level of probability is used, corresponding to a value of $t = 1.96$. In this chapter, the value of 2 is used as an approximation to 1.96.

Our aim is to reduce the length of the confidence interval without decreasing the probability of making truthful statements. The reduction of the standard error of sample results is the goal of sample design, discussed in Sections 7 and 14.

3 Confidence Intervals

With regard to the use of confidence intervals for sample estimates, some understanding of several important points is necessary

(a) From the results of well-planned probability samples, it is possible to compute values for the estimated standard errors of sample means. Mathematical statisticians have derived formulas from which we can compute those values or some useful practical approximations. The standard error is defined as the standard deviation of the distribution of the sample means, and the distribution depends on the specific sample design used. Therefore, the formula for the estimated standard error will depend on the sample design used. In the case of simple random sampling, we have the familiar forms of s / \sqrt{n} or $\sqrt{p(1-p) / (n-1)}$. But these formulas will not hold for other sample designs. The difference may be either in the selection or in the estimation procedures.

(b) Many of the estimates used in practice have sampling distributions which are approximately normally distributed. Just how good that approximation is depends on the underlying distribution of the characteristic in the population and on the size and design of the sample. For any variable encountered in practice, the approximation improves with the size of the sample.

In case one has a small sample, or some other reason for doubts, he should ascertain whether he may proceed with the assumption of normality. If the sampling distribution departs seriously from the normal distribution, two kinds of alternatives are open for the construction of confidence intervals. A search may be made for some distribution other than the normal, to serve as a useful approximation. Or he may try to make use of a "distribution free" statistic (see Chap. 12).

For many sample results encountered in practical social research work, the assumption of normality will lead to errors which are small compared to other sources of inaccuracy which are tolerated. The assumption of normality leads to simple statements of probability through the use of confidence intervals. If, as below, we use tables of the normal distribution for making those probability statements, we thereby assume that the normal distribution is a good approximation to the sampling distribution of our estimates. In so far as that assumption is not justified, our statements will have a probability of being wrong different from that which we intended and stated.

congruent with the probability model that underlies our statistical theory. Terms such as "tossing perfect coins" or "drawing perfect balls from perfect urns after complete mixing"—terms found in textbooks—have the aim of providing an intuitive grasp of the kind of necessary physical process.

However, there are practical reasons why we must make two modifications in this simple picture. First, there are serious objections to tossing, mixing, and "drawing" the kind of elements with which we want to deal—the elements themselves might object. This problem is met by listing and numbering the elements, and then mixing and drawing uniform objects bearing their identification numbers. That is what is supposed to go on in a bingo cage, or in a lottery. Secondly, it is difficult to construct perfect coins, perfect balls, or perfect urns, or to bring about complete mixing. The equivalent of that process has been performed by careful experts who gave us their results in "tables of random numbers"—our convenient equivalent of complete mixing of perfect balls. Thus, we see that the mechanical process of selection, which is indispensable for probability sampling, is accomplished by the following chain. A set of numbers selected properly from a table of random numbers identifies a set of numbers on a listing of individual sampling units, from these selected units the identification is made to a set of physical units which will comprise the sample.

Sometimes there are practical problems involved in the identification of the individuals associated with the selected numbers on the listing. These are problems of field procedure which must be solved with clear, simple, and practical instructions, and not merely assumed. Sometimes, as in the case of identifying employees from a payroll list, the task may be simple and easy. At other times, as in identifying dwelling units from a block listing sheet, the process may call for skill, and mistakes may occur (see Section 15).

Furthermore, there are some criteria which a list must fulfill, and the making of a satisfactory list involves various problems. First, a list may be incomplete, for example, the list of payroll cards of the factory may exclude white-collar employees or the employees hired since some recent date. It may be decided to exclude these explicitly from the population. Alternately, one may establish a separate stratum for them so that they will obtain the proper probability of selection through separate sampling procedures.

4. *Measurability The Need for Probability Samples*

The use of confidence intervals is based on statistical theory. The application of this theory can be developed only for those samples in which the probability of selection of every element of the population is known. There is a gulf between the known sample result and the unknown population value, the confidence interval is the only objective statistical bridge across that gulf. Confidence intervals are based on the proper estimates of the standard error. But standard errors may be calculated validly only for probability samples—that is, for samples for which the probability of selection of every element in the population is known.* Therefore, if we wish to make use of the theory of statistical inference, we must use a probability sample.

The property of a sample which enables the researcher to make estimates based on sample data of population values and then to calculate confidence intervals for those estimates has been called measurability. It is desirable to consult a sampling statistician with the plans *before* the survey to see whether the design will allow the valid calculation of the precision of the sample. There may be other ways of judging the adequacy of samples, but they depend on personal judgment. If a nonprobability sample is taken, such as a "quota sample" or a "typical" city, the results may be good or they may be poor. But statistical theory is lacking for determining the accuracy of the results. There may be occasions when for a small informal sample one can afford to dispense both with precision and with its measure. However, this chapter will be confined to the problems of probability sampling.

5. *Mechanical Selection and the Use of Listings*

How does one select a probability sample? Whenever a unit is selected into the sample from among other units, the selection must be made by some mechanical procedure which guarantees the desired probabilities of selection to all the units involved. We need some physical operation, some practical procedure, which will be reasonably

*Recently the term *random sampling* has been used by some authors synonymously with *probability sampling*. In this chapter, when the phrase *random choice* is used, we shall mean a process of selection with equal probability among the defined group of sampling units.

in the various phases of area sampling, presented in Sections 19, 20, and 21. Of particular influence is the size of the sampling unit used in the listing which results in clustering, subsampling, and multistage sampling (3, pp 76-87, 22, pp 60-80)

Area sampling is an important kind of listing procedure because it is used widely in social studies. It is also used in other types of surveys; for example, crops or other flora, as well as grocery stores, have been sampled with the use of area segments. Its widespread use in social surveys is due chiefly to the relative ease of identifying each member of a human population with one, and only one, dwelling unit. In turn, these dwelling units are identified with area segments, also uniquely. Thus, a selection of the area segments yields a sample of dwellings, and these in turn a sample of people. It may be expected that the unique identifications of people with dwellings, and of these with segments, is troublesome and imperfect. This becomes a practical matter of doing a good job within available resources. In this connection one may mention the necessity for boundaries which are clear, unambiguous, and easily identified in the field.

7 Precision, Variations in Sample Design

It is our general goal to obtain as small a confidence interval as we can for some fixed level of probability of making correct statements. The smaller its confidence interval, the more useful is the sample estimate. For a fixed probability level, the length of the confidence interval depends on the standard error. For this reason the word *precision* is often used to denote the inverse of the standard error (or sometimes of its square, the variance) of the sample estimate.

The standard error of a simple random sample is s/\sqrt{n} . Within the limits of that design, the way to increase the precision (to reduce the standard error) is to increase n , the number of elements in the sample. The standard errors of other sample designs are different, but they all have the common property that to increase the precision we must take more of something (persons, dwellings, blocks, counties, etc.). But to take more of anything costs money—and generally there is a limit to what the sample may cost. The question may be asked, then: For a given expenditure, how can we get the greatest precision? (See also Section 14.)

Secondly, some elements may appear more than once on the list. Perhaps employees who work in more than one department will have more than one payroll card. If these individuals are to have the same probability of selection as the others then one should eliminate the duplications through the entire list. If this would be too difficult, some other adjustment has to be made in our procedures.

Thirdly, a list may contain other items in addition to members of the population. Some lines may be simply blank, belonging to no units, and others may belong to units of some other and excluded population. Let us say that our list identifies the 12,000 employees of the factory by their actual payroll numbers. These are not consecutive, but run irregularly up to 99,999. Thus, among the five-digit numbers, there are 88,000 which do not belong to any of the 12,000 members of our population. Some may be blank and some may belong to office employees, whom we want to exclude from the sample. We simply draw five digit numbers from a table of random numbers. We inspect the list for each drawn number. If it belongs to a member of our population, we have a selection, otherwise, we have not. We continue this process until the designated size of the sample (n) is reached. This gives us a simple random sample of n elements out of the designated population. In general, the list should be inspected carefully before drawing to take whatever measures are necessary for the insurance of the proper probabilities of selection.

6 *Listings and Area Sampling*

The nature of the lists available for the selection process is an important consideration in the design of the sample. Factors which are relevant include the nature of the listed sampling units, the extent of coverage, the accuracy and completeness of the list, and the amount of auxiliary information on the list. This last factor is useful, as we shall see, for stratification, for measures of size, and in the estimation process. Those factors help to determine the nature of the sampling design and of the details of the practical selection procedures.

In many sampling situations there does not exist a simple and complete listing of all individuals, such as the factory payroll discussed in Section 5. Moreover, it may be too costly to construct one. The important practical complications which a list may possess are present

Attached to many formulas in textbooks is an important but often overlooked phrase which reads something like this "Given n independent selections made with equal probability" Actually such a selection is seldom given to the researcher Furthermore, in practical social research such a selection is seldom taken by the researcher Therefore, the automatic use of those formulas which assume simple random sampling is not justified Sometimes their unjustified use leads to the construction of confidence intervals which are much too narrow The result in those cases is that the estimate will lead to incorrect statements considerably more often than the researcher would wish (Section 13) Serious mistakes of this kind are frequently made, owing to the disregard of the clustering present in the selection process

The proper calculation of the variance of a sample estimate must be done in accord with the procedures used in the sample design In order to call attention to that need, a specific subscript was used to denote each sample design Thus \bar{x}_o' denotes a sample mean obtained from a simple random sample (Section 1), \bar{x}'_{prop} denotes the mean of a proportionate stratified sample (Section 9), and so on

One word more of caution may be appropriate here What may appear to be a slight change in the sample design may affect seriously the variance of the sample estimate Furthermore, some change in procedure may introduce a serious bias into the design

STRATIFICATION TECHNIQUES

8 Stratification

Stratified sampling is the procedure of dividing the population into subpopulations, called *strata*, and then selecting a sample within each Every sampling unit in the population is placed in one (and only one) of the strata prior to the selection of the sample so that the sum of the strata is identical with the population Within each stratum, a sample is selected from the units in that stratum, and from each of those samples the estimate is calculated for its stratum Finally, the separate estimates for each stratum are combined to form an estimate of the total population value (However, the simpler calculation is available for the means of proportionate stratified samples, given in Section 9)

Now we calculate

$$\bar{x}_w' = 6 \times 20 + 3 \times 35 + 1 \times 45 = 27$$

That is, 27 percent is the combined sample estimate for the proportion who would have answered "yes" to that question if all the employees of the company had been questioned

The estimated variance of that sample result is

$$\text{est var } (\bar{x}_w') = (6)^2 (0.20)^2 + (3)^2 (0.35)^2 + (1)^2 (0.40)^2 = 0.00196$$

That is, the estimated standard error is

$$\text{est se } (\bar{x}_w') = \sqrt{0.00196} = 0.04$$

Therefore, with the use of two standard errors, the statement is made that the percentage of the entire 20,000 employees of the company who would have said "yes" to that question is between the limits $27 \pm 2(0.04)$, that is, between 24.2 percent and 29.8 percent. This statement has a probability of about 0.95 of being correct.

Note the important implications of the above. From each of three subpopulations a sample was chosen *separately*. The designs, the procedures, and the sizes of the samples in each of the strata are unknown and irrelevant. We have the sample means and their standard errors as they were calculated separately from the three samples. Now it is desired to make an estimate for the combined population represented by the sum of the three subpopulations (strata). The relative number of elements in each of the three strata is the weight of that stratum, with the use of these weights, the separate stratum estimates are readily combined to obtain estimates for the entire population.

There are three kinds of reasons for using a stratified sample.

(a) Stratification may be aimed at the *reduction of the variances* of the sample results for the entire population. Thereby greater precision is obtained for the sample estimates, a constant goal of sample design. Section 9 discusses one method of aiming at that goal, Section 11 discusses contrasting methods (1, pp. 65-110).

(b) It may be thought convenient or necessary to use *different sampling methods* or procedures in different strata of the population. For example, in one of the factories of the illustration, a sample of individuals may be selected, in another, the sample may be selected in

The process of selection in each stratum is carried out separately and independently. In each of the strata one may use a different sampling fraction, and even different methods and procedures. Regardless of the procedures used within the different strata, the strata means may be combined to form an estimate for the population thus

$$\bar{x}_w' = \sum_{h=1}^R w_h \bar{x}_h'$$

That is the sample mean \bar{x}_h' obtained separately for each stratum is multiplied by the weight of that stratum, then these products are summed over the R strata to obtain the combined weighted estimate \bar{x}_w' for the entire population. The weight w_h of each stratum is the proportion of the total population contained in the stratum. The sum of these weights is 1 that is, $\sum_{h=1}^R w_h = 1$.

The variance of the combined weighted mean will be the sum of the variances of the individual strata means, each multiplied by the square of the stratum weight

$$\text{est var } (\bar{x}_w') = \sum_{h=1}^R w_h^2 [\text{est var } (\bar{x}_h')]$$

As an example imagine that the factory mentioned in Section 2 is only one of three belonging to a company and that a separate survey of attitudes was conducted in each. One of the questions appeared in each of the three surveys and it is now desired to estimate what percent of all employees of the company would have said "yes" to this question

Stratum (factory) number (h)	1	2	3	Company total
Number of employees in stratum (N_h)	12 000	6 000	2 000	20 000
Relative proportion of employees ($w_h = N_h/V$)	0.6	0.3	0.1	0
Sample mean (the proportion of yes answers) for stratum (\bar{x}_h')	20	35	45	
Estimated standard error of sample mean of stratum [$\text{est se } (\bar{x}_h')$]	0.20	0.20	0.40	

However, the information used for the sorting of units into strata need be neither objective, accurate, nor complete, on the contrary, personal judgment can often be used profitably for this purpose

The role of stratification in sample design is often misunderstood and exaggerated. Sometimes it is implied loosely that if stratification is used we may shut our eyes to other aspects of the selection procedure. This implication is used to justify the use of nonprobability methods of selection, such as "quota" samples. This attempt to establish stratification as a sufficient condition for an adequate sample design has no basis in statistical theory. Each stratum is a subpopulation, and the principles of probability sampling must be applied to the selection procedures used within the strata.

Sometimes it is implied that stratification is a necessary condition, a "must," for an adequate sample design. It is far from that. Actually the sampler usually has to be satisfied with relatively dull variables (such as age, sex, etc.) because the more interesting variables (such as basic psychological make up or social history of the individual) are not available. In many practical cases the gain from stratification is little, that is, the same precision may be achieved with but a little more expense without the use of stratification. Nevertheless, stratification is used in most sampling undertakings because it is generally beneficial and because it is easy to apply.

9 A Proportionate Sample of Elements

Proportionate stratified sampling of elements is often in the back of people's minds when they talk of "representative sampling," when they insist that the "different parts of the population must be properly represented." Let us describe a *proportionate stratified random selection of elements* by means of an example. Suppose that we want a sample of $n = 400$ employees out of the $N = 12,000$ employed in a factory. We suspect that there may be important differences in the attitudes to be measured among the employees in the different departments. Therefore, the employees are listed separately for each of the four strata (departments) into which the company was divided.

In order to have a *stratified* sample, the selection must be carried out separately in each stratum. It is a selection of *elements* because the elements (employees) are selected individually, separately. In order to make it a *proportionate* sample, the number of elements from each

clusters of work sections. These differences in procedures may be in accord with differences in the physical distribution of the population elements or with differences in the manner in which they are listed, or they may be due to differences in total survey objectives within the different strata.

(c) Sometimes strata are established because the subpopulations are also designated "*domains*" of study—that is, the survey is designed to provide sample results of some desired precision about the several subpopulations separately, as well as about the combined whole. In our illustration, the sample results (\bar{x}_k') were desired for each of the three factories, the domains of that study, and the desired precision (s.e. of \bar{x}_k') was used to determine the design and the size of the sample to be taken in each.

What characteristics of the sampling units of a population should be used for stratifying the units in order to reduce the variances of the sample estimates for that population?

(a) The stratifying characteristics should be related to the variables to be estimated from the study. The sorting of the sampling units on the basis of the stratifying characteristics should establish strata which will turn out to be (relatively) well sorted with regard to the variables to be studied. The reduction of the variances of the sample results is achieved in so far as the variation (of the characteristics studied) among the sampling units within the strata is less than their variation throughout the population. Hence, in stratifying, one attempts to make the sampling units within the stratum as homogeneous as possible.

(b) In most cases it is wasteful to spend much time worrying about just which variables would be most suitable. Experience shows that usually there is not much difference in the precision of two procedures if both are based on some reasonably good stratifying variables. A person acquainted with the subject matter will usually hit readily on a reasonable choice. An elaborate and expensive further search may not bring commensurate gains.

(c) Each sampling unit must be assigned to one of the strata.[†]

[†]If the information is not available for some of the sampling units, a "miscellaneous stratum" is devised. Sometimes "double sampling" is used to obtain information cheaply in the first phase for stratifying a large sample, then, in the second phase, a smaller sample is subselected for the main part of the study (1, p. 268, 22, p. 38, 16, p. 153).

	Assembly	Foundry and machines	Office	Other	Entire factory
Stratum number h	1	2	3	4	Total
Stratum weight w_h	333	250	250	167	1 00
Number of employees selected from stratum n_h	133	100	100	67	400
Number of "yes" answers in stratum X_h'	12	11	36	21	80
Proportion of "yes" answers					
$\frac{1}{n_h}X_h' = \bar{x}_h'$	09	11	36	31	

Because the sample is self-weighting, the mean for the sample may be taken simply as the total for the sample divided by the number of cases in the sample ¹

$$\bar{x}_{prop}' = \frac{1}{n} \sum x$$

In our example, we have simply $80/400 = .20$. The subscript "prop" of the sample mean identifies the sample design with which it was obtained. The term *self-weighting* denotes that, in calculating the sample mean, the sample cases are simply added without any special weighting procedure.

If we had a stratified sample which was not proportionate, the sample would not be self-weighting (Sections 11 and 22). In that case the mean would have to be obtained by a weighting procedure.

$$\bar{x}_w' = \sum w_h \bar{x}_h' =$$

$$333 \times .09 + 250 \times .11 + 250 \times .36 + 167 \times .31 = 0.20.$$

¹As given in Section 8, $\bar{x}_w' = \sum u_h \bar{x}_h'$. The mean for the stratum is $\bar{x}_h' = \frac{1}{n_h} \sum x_h$.

In a proportionate sample, $u_h = n_h/n$. Therefore, $\bar{x}_{prop}' = \sum \frac{n_h}{n} \frac{1}{n_h} \sum x_h = \frac{1}{n} \sum \sum x_h$. But the double summation means simply that after the sum of all values in each stratum is obtained, these partial sums must be added for all R strata to obtain the sample total. This summation can be done all in one step and we may express it as $\bar{x}_{prop}' = \frac{1}{n} \sum x$.

		Foundry and Assembly machine				Entire factory
	Symbol	Office	Others			
Stratum number	h	1	2	3	4	Total
Number of employees	N_h	4,000	3,000	3,000	2,000	12,000
Relative weight (N_h/N)	w_h	333	250	250	.167	1 000
Number of employees to be selected (nu_h)	n_h	133	100	100	67	400

stratum in the sample must be made proportionate to the number of elements from each stratum in the population. The numbers of elements in each stratum *relative* to the population total (N) is denoted by the stratum weight $w_h = N_h/N$. Now if we multiply the total desired size of the sample (n) by the weight of the stratum, we obtain the number of elements to be selected in each of the strata $n_h = nw_h$ (as shown on the last line). Thus the sample will be proportionate because the representation of each stratum in the sample is equal to the ratio of that stratum in the population $n_h/n = N_h/N$. For example, $133/400 = 4000/12,000 = .333$.

Another way of regarding a proportionate sample is that the sampling fraction in each stratum is equal to the sampling fraction for the population as a whole $n_h/N_h = n/N$. That is, the sampling fraction is $1/30 = 400/12,000 = 133/4000 = 100/3000 = 67/2000$. (There is a small difference due to the unprecise fraction in this case, as there is usually in real cases. This is ordinarily a trivial matter.) That is, the sampling fraction $n/N = 400/12,000 = 1/30$ is obtained, then this fraction is applied to the numbers of elements (N_h) in each of the strata.

There is one more word to be explained in our definition of the sample design *random*. By random we shall mean that the selection of the n_h elements out of the N_h in each stratum is to be made by a separate random choice with equal probabilities among all the elements, just as defined in Section 1. In stratum 1, for example, 133 different random numbers from 1 to 4000 must be taken from a table of random numbers. That is, *within* each stratum we select a simple random sample.[†]

[†]A systematic selection taking every 30th employee is described in the next section. This is more usual in practice.

formula belongs to simple random sampling (Section 1), a sampling method different from the one which was actually used. In the present case the difference is not great, but in Section 13 some great differences due to clustering will be noted.

The ratio $.000354/.000388 = 91$ percent is a measure of the 9-percent gain in precision over simple random sampling which is due to the stratification in this case. That is, it would take about $400/.91 = 440$ interviews by simple random sampling to get the same variance as the 400 interviews in the proportionate stratified random sample of elements. Several things may be said about proportionate sampling of elements:

(a) The gains in precision arise because each stratum is proportionately represented in the sample. This helps by eliminating from the variance of the mean of the combined sample that component of the variation which is due to differences among the means of the several strata. Another way of looking at this is to note that in this type of sampling the variability arises only from sampling within the strata. In so far as the strata are homogeneous—that is, in so far as the variability within the strata is less than in the population at large—just so far will the stratification be useful.

(b) The gains from the use of proportionate sampling of elements are usually not spectacular. In the example the gain was 9 percent, which may be looked on as a 9-percent reduction in variance for a given number of interviews (about $4\frac{1}{2}$ -percent reduction in the standard error); or as a 9-percent saving in the number of interviews needed for a fixed variance (see Section 14). Actually, this saving is greater than one usually finds in practice and it is due to the remarkable differences of the proportions (p_h) in the four strata. Let us say, for example, that the proportions holding for another attitude on the same survey are 40 percent, 50 percent, 50 percent, and 60 percent in the four strata; the gain from stratification for that characteristic is only 2 percent, the equivalent of 8 interviews in 400. The gains from stratification may be shown by simple formulas (3, p. 227; 8, p. 11). This modest gain is in sharp contrast with rather widespread, vague, notions of stratification. Perhaps some understanding may be gained from the consideration that without any stratification a simple random sample would obtain in most cases *nearly* the correct proportions from each stratum. In our example, a proportionate sample allocates 133 out of the 400 interviews to the first stratum. Without any stratification a

simple random sample would obtain some number between 124 and 142 two thirds of the time, and between 114 and 152 in 19 samples out of 20

(c) We shall see in Section 11 that in some cases bigger gains may be obtained from allocations of the sample which are not proportionate. And in Section 22 an example is given to show that the gains of proportionate sampling may be obtained without stratified selection, by weighting the results in the separate strata so that in the combined sample result the various strata are properly represented

(d) Although the gains from it may be small, proportionate stratified sampling is very widely used. One reason is that it is a safe thing to do: the precision cannot be worse than if the sample is drawn without stratification, and often it is better. Secondly, it is an easy thing to do: quite often, as in our factory example, it can be done with little or no effort. Thirdly, because the sample is self-weighting, the calculations are simpler than with the use of either of the two methods mentioned in the preceding paragraph.

(e) The gains arising from stratification are often greater when the sampling units are clusters than for the selection of elements. Examples of stratified samples of clusters are discussed in Sections 18-21.

(f) For a proportionate sample of elements, one should not waste much time considering the exact variables to use for stratification. In most cases the researcher can choose readily the useful stratifying variables, in so far as they are available. In our example of the factory, the departments suggested themselves. Perhaps sex, age, work sections, or job classification could have been used instead of the departments or as additional strata.

10 *A Systematic Sample*

In the taking of a sample of the employees of a factory (students of a school or members of a club, etc.) a proportionate sample of individuals would be a likely choice for sample design. Let us assume that the payroll has been listed by departments and that within the departments the names are arranged in a haphazard order or perhaps alphabetically. We calculate that we want every employee to have a $400 - 12,000 = 1$ in 30 chance of selection. In Section 9 we went on to designate the numbers $n_k = n(N_k/N)$ that were to be selected in each

department, then from the N_h employees in the department n_h random numbers were selected

In practice a simpler device is commonly used systematic sampling This is probably the most widely known design, it consists of taking every k th individual after a random start from 1 to k In this instance one would take a number from 1 to 30 from a table of random numbers With that number as a start the interval of 30 is applied If the random start was number 28, we should have the numbers 28, 58, 88, 118 11,998 in the sample These numbers refer to the consecutive numbers of the employees as they appear on the list The estimate of the mean here, as in a proportionate sample, is the usual simple mean of the sample $\bar{x}' = \frac{1}{n} \sum x$ However, the estimation of

the variance of the mean is not clean cut (1, pp 179-182)

If the payroll cards of each department had been shuffled thoroughly before they were ordered on the list, the systematic sample would be equivalent to a proportionate stratified random sample For the latter, the shuffling process is not necessary because the procedure of n_h separate independent choices in each stratum accomplishes the equivalent of a shuffling process With the regular intervals of systematic selection this shuffling is lacking However, in many practical instances the haphazard arrangements are considered to give results similar to random selection within the strata, in those cases the formula of stratified sampling would be used

$$\text{est var } (\bar{x}_w') = \sum w_h^2 \frac{s_h^2}{n_h}.$$

Nevertheless, we should be wary In stratified random selection the arrangements of the units within the strata may be ignored because the random selection will provide the necessary shuffling With systematic sampling there is need for reasonable reassurance that the arrangements of the sampling units within strata may be regarded as if they were random There are schemes for using several different random starts rather than just one

The researcher should be alert for two kinds of departures from randomness in the arrangement of the population units If either of these situations exists, or may exist, a systematic sample should be avoided or modified

(a) *A trend* Imagine that, unknown to us, somebody sorted all the 3000 employees in the foundry according to increasing seniority, and that is the order in which they appear on the list from which we shall select the sample. Then we are to select our systematic sample of every 30th after a random start. One of the 30 possible samples would select the employees numbered 1, 31, 61, ..., 2971, another sample would consist of numbers 30, 60, 90, ..., 3000. In the latter possible sample each employee has more seniority than his counterpart in the former by 29 ranks. The means of these two samples would be widely divergent as regards seniority and other variables strongly associated with it. Hence, there is a great deal of variability among the 30 possible samples, results on any characteristic strongly related to seniority would be dependent on which of the 30 possible samples happened to be chosen.

(b) *A cyclical fluctuation* Imagine that one of the departments is composed of work sections each of which contains 10 employees and that somebody has arranged the listing so that the 10 employees of each section are together and in the order of their seniority. Thus we see a cyclical fluctuation of seniority with a "period" of 10 employees. Now if a sample of every 30th employee is selected, there are 30 possible samples. Three of these samples (those with the random starts of 1, 11, or 21) would select the most senior employee in each group of 10, three others the most junior, and the other samples would fluctuate similarly. Again, we see a great deal of variability depending on which among the 30 possible samples happens to be selected (1, pp. 160-174).

In addition to an unduly great amount of variability, there is another serious objection to a systematic sample under these circumstances. The calculations made from the sample will not show this source of variability. Thus the true standard error may be grossly underestimated from the sample data.

Nevertheless, systematic sampling is used very widely. Because of the irregular arrangement of so many kinds of populations, the formula for stratified random sampling is frequently an adequate approximation, as we noted above. However, suppose that within each department of our illustration there is an arrangement (by work sections, for example) which makes for stratification within each department. That stratification would not be reflected in the formula based on the assumption of simple random selection within the department. A

formula which will be a useful approximation in these cases is one which is based on the successive squared differences among the n selected elements in their order of selection (1, p. 180; 22, p. 229):

$$\text{est. var. } (\bar{x}_{\text{sys}}') = (1 - f) \frac{1}{2n(n-1)} \sum^{n-1} (x_i - x_{i+1})^2.$$

In Sections 12, 19, 20, and 21, examples of systematic selections of clusters are given. In Section 18, a formula similar to the one above is discussed in application to a cluster selection.

11. Allocation to Strata

A method of using stratification to increase the precision of the sample mean \bar{x}_w' is the deliberate use of different sampling rates in the various strata. The estimate $\bar{x}_w' = \sum \frac{N_h}{N} \bar{x}_h'$ can be made most precise for a fixed cost if the sampling rate within each stratum is made directly proportional to the standard deviation within the stratum and inversely proportional to the square root of the cost per element in that stratum. That is, for a minimum standard error of \bar{x}_w' , often called "optimum allocation," make $\frac{n_h}{N_h}$ proportional to $\frac{s_h}{\sqrt{J_h}}$.

Here J_h is the cost per element, so that $\sum n_h J_h$ is the fixed total cost related to the number of interviews (1, p. 73).

Several points may be noted:

(a) If the cost per interview is the same in each stratum, the problem becomes one of allocation of a fixed number (n) of interviews to the various strata. In that case the sampling rate in each stratum should be made proportional to standard deviation within the stratum.

(b) This "optimum allocation" of the sample may also be viewed as that which yields a desired variance of the estimated mean for the least cost (see Section 14).

(c) Of s_h and J_h only rough estimates are usually available. However, precision is not needed here; convenient rates which are roughly proportional to those quantities ordinarily suffice. The difference in precision is small between an optimum allocation and another which is only roughly like it (3, p. 366).

(d) For the reason just given, ordinarily it does not pay to resort to disproportionate allocation unless there are substantial differences among the estimates for s_h or those for J_h among the various strata. If those differences are large, the gain over proportionate sampling may be large. However, in the estimation of proportions, usually no great gain may be had through disproportionate sampling, their standard deviations are $\sqrt{p_h(1-p_h)}$, and that quantity is not sensitive to the kind of fluctuations one usually encounters with values of p_h between 10 and 90.

(e) Disproportionate allocation should be used only with caution, perhaps only on expert advice. The optimum allocation for one item on a survey may result in large losses of precision (greater standard errors) for some other items on the survey. Furthermore, although the "self weighted" proportionate sample makes for easy computations, the calculation of the sample mean of a disproportionate sample involves weighting in inverse proportions to the sampling rates. The weighted calculations of a disproportionate sample may be costly. This added cost of tabulation, not included in the formulation of "optimum allocation" above, should be considered before a sample design with disproportionate sampling rates is adopted.

CLUSTERING

12 *Cluster Sampling*

In a factory many instances would arise that would call for a cluster sample. It might happen, for example, that employees would be selected not individually but in clusters of work sections. Let us move on, however, to another illustration. Let us say that we want a sample of 400 out of the estimated 12,000 dwellings of a city, with equal probability of selection for each dwelling. Were a list of the city's dwellings available, the procedures of sampling discussed in Sections 1, 10, and 11 could be used here, too. Suppose, however, that such a list is not available and it is deemed too costly to prepare. Suppose, furthermore, that it is desired to economize on the costs of locating dwellings by means of sampling entire blocks. The entire area of the city's map is divided into blocks along identifiable streets and other

natural boundaries so that every dwelling is located clearly inside one of the blocks. The blocks are numbered consecutively, this numbering establishes a list of the blocks, the numbers going from 1 to 750. The number of dwellings per block is variable, but the average is 12,000 — $750 = 16$. Now a sample is chosen by selecting in a random manner 25 blocks out of 750 and including in the sample all dwelling units found within the boundaries of each of the 25 sample blocks. Note that the probability of selection of any dwelling in the city is $25/750 = 1/30$.

In this example the elements are dwelling units, since the analysis will be in terms of characteristics of dwelling units. The sampling unit, however, is the block, the selection is made from a complete list of all the sampling units in the population—that is, the numbered list of blocks. Each element belongs to one, and only one, of these sampling units. The selection of each sampling unit results in the selection of the cluster of elements which it contains.

Cluster sampling is the name given to methods of selection in which the sampling unit, the unit of selection, contains more than one population element, the sampling unit is a cluster of elements. In our illustration the block is the cluster, composed of dwelling units as elements. But just what is a cluster or an element is only a matter of practical expediency. In some studies the dwelling unit will be regarded as a cluster of persons, whereas in another study the population elements might be blocks or even cities. The same physical population, the people of the United States, may be regarded in turn as composed of units which are states or counties or cities, towns, and townships, or blocks or dwellings or, finally, the individual persons. The elements of the population are defined in accord with study objectives. The clusters are defined in conformity with the requirements of a practical and economical sampling design applied to the physical distribution of the population (3, pp 135-146).

After the individual elements are defined, the question arises whether they should also serve as sampling units or whether, contrariwise, it may be more economical and practical to define a sampling unit which is a cluster of the previously defined elements. There will be several examples given, in later sections, of situations which lead to the use of clusters. The choice of clusters is generally a recognition in the sampling procedures of some existing features in the make-up of the population. Thus, one might study the employees of a factory

and select work groups, one might study the students of a university and select them in clusters of classes, one might select dwellings in clusters of blocks. In any case, one must ensure that every element of the designated population belongs to one, and only one, of the clusters. Otherwise, special measures must be taken.

CLUSTERS OF EQUAL SIZE Clusters of equal size rarely exist in society. They come into being only as the result of planning. Army units, the units of large housing developments, and some work sections in large establishments may sometimes be of equal size, or nearly so. Sometimes, however, the sampler creates equal-sized clusters where none existed before, as in the example given below. Furthermore, it is common practice to create equal sized subsample clusters by the procedures of sampling unequal sized clusters with probabilities proportional to size discussed in Section 20. Clusters of equal size can be treated simply as a special case of samples of unequal size. However, the subject of clusters of unequal sizes is rather complex.

As an example, take the file of subscribers of a newspaper. There are 12,000 subscribers served by carrier routes. There is a card for each subscriber in the file. The 100 to 200 cards of each carrier route are kept together, and neighboring routes follow one another. An interview survey of about 400 subscribers is wanted, and in order to save travel time it is decided to take clusters of 10 subscribers each. It is estimated that 10 short interviews in the same neighborhood can be generally obtained in half a day's work on the spot.

Now, imagine that the file is divided into 1200 clusters of 10 consecutive cards each. A sample of 40 of the 1200 clusters is to be selected. The drawing of 40 different random numbers from 1 to 1200 would give a random selection of the clusters.* However, our aim is to be practical, and in practice a systematic sample of clusters would generally be taken (see Sections 10 and 18). Let us say that after a random start from 1 to 30, every 30th cluster was chosen, we thus have 40 clusters of 10 cards each. Altogether 400 out of 12,000 were selected, each with a probability of selection of 1 in 30.

To divide the file into 1200 clusters would take some time, so

*Note that random selection here denotes 40 independent selections with equal probability from the 1200 units in the population. The sample of 400 cases is obtained by means of only 40 choices. This has important consequences for the sampling error, as discussed in the next section.

would the counting of 290 nonselected cards for each cluster of 10 selections. In practice, some approximate procedure might be used, such as the use of a ruler to measure off the clusters. Let us say that of the 400 subscribers interviewed 148 gave a "yes" answer to a question. The number of "yes" answers out of the 10 interviews in each of the 40 clusters were also obtained. These are, with the sections arranged in the same order in which they were selected:

4 7 1 1 3 5 3 6 2 3 2 2 1 4 4 6 3 2 7 1
5 5 4 6 8 2 0 3 5 6 9 2 1 0 3 4 7 5 2 4

The sample mean is calculated as before by dividing the sample total by the number of cases:

$$\bar{x}_{ec}' = \frac{1}{n} \sum x.$$

In the present instance we have $\bar{x}_{ec}' = \frac{1}{400} 148 = .37$.

Denote by m the number of clusters and by A the number of elements per cluster, so that $mA = n$. (In the present instance we have $40 \times 10 = 400$). Let X_i denote the sum of the values of the characteristic x for the A values in the i th cluster; in our example, the 40 values of X_i are given above as 4, 7, 1 . . . 5, 2, 4. The mean of the i th cluster will be denoted as $\bar{x}_i = \frac{1}{A} X_i$. These 40 values may be found by dividing the values above by $A = 10$; hence they will be 0.4, 0.7, 0.1, . . . 0.5, 0.2, 0.4. These represent the mean of the characteristic for the elements in the cluster—that is, the proportion of subscribers in the different clusters who said "yes." The subscript "ec" is used here to denote "equal clusters," whereas the subscript "c" is used later when the clusters are not necessarily equal. It may be noted that the sample mean is also equal to the mean of the cluster means:

$$\bar{x}_{ec}' = \frac{1}{m} \sum \bar{x}_i.$$

In our example, $\bar{x}_{ec}' = \frac{1}{40} (.4 + .7 + .1 + \dots + .2 + .4) = .37$.

In order to obtain the sample mean, it is not necessary to go to the

trouble of calculating the separate cluster values. However, we need them to obtain the estimated variance of the sample mean

$$\text{est var } (\bar{x}_{ec}') = (1 - f) \frac{s_b^2}{m}$$

Here $s_b^2 = \frac{1}{m-1} \sum (\bar{x}_i - \bar{x}_{ec}')^2$, the variance of the cluster means around the sample mean. Note the similarity of the formula to that for a simple random sample: here, too, we have a variance divided by the *number of independent sampling units*. The units involved in the variance calculations are also the sampling units: the clusters. The quantity $(1 - f)$ is again the usually inconsequential "correction for finite population." With m clusters selected out of a total of M in the population, m/M is the sampling fraction, and $(1 - f) = 1 - m/M$. In the present instance we have $(1 - f) = 1 - 40/1200 = 1 - 1/30$.

We have †

$$\begin{aligned} \text{est var } (\bar{x}_{ec}') &= \frac{29}{30} \frac{1}{40} \left(\frac{1}{39} \times 1.964 \right) = \frac{29}{30} \frac{1}{40} \times .0504 = \\ &967 \times .001259 = .001217 \end{aligned}$$

$$\begin{aligned} \text{Here } \sum (\bar{x}_i - \bar{x}_{ec})^2 &= (4 - .37)^2 + (7 - .37)^2 + \quad + (2 - .37)^2 \\ &\quad + (4 - .37)^2 = 1.964 \end{aligned}$$

The standard error is $\sqrt{.001217} = .035$, and the confidence intervals are $.37 \pm 2(.035)$. That is, we make the statement that the proportion of "yes" answers that would have been obtained by a similar survey

†More convenient calculational forms are

$$\begin{aligned} \text{est var } (\bar{x}_{ec}) &= (1 - f) \frac{1}{m} \left[\frac{1}{m-1} (\sum \bar{x}_i^2 - m \bar{x}_{ec}^2) \right] \\ &= (1 - f) \frac{1}{m(m-1)A^2} \left[\sum X_i^2 - \frac{1}{m} (\sum X_i)^2 \right] \end{aligned}$$

$$\text{This last form comes to } \frac{29}{30} \frac{1}{40 \times 39 \times 100} \left[744 - \frac{1}{40} (148)^2 \right]$$

$$\text{Here } \sum X_i^2 = 4^2 + 7^2 + 1^2 + 2^2 + 4^2 = 744$$

And $\frac{1}{39} \cdot \frac{1}{100} \left[744 - \frac{1}{40} (148)^2 \right] = .0504$. This quantity, as above, is the value of s_b^2 , the estimate of the variance of single clusters in the population.

of the entire population of 12,000 subscribers is between 30.0 and 44.0 percent. That statement has a 95-percent chance of being correct and a 5-percent chance of being wrong. It should be noted that this formula for the variance is entirely appropriate only if the m clusters are made with m random choices. But for practical reasons the selection was systematic. The results of our systematic selection would be equivalent to a random choice only if the order of the clusters in the file had been thoroughly randomized. They were not, and we are told that "neighboring routes follow one another." This matter will be treated in Section 18, where we shall find that for our example the approximation is not bad.

CLUSTERS OF UNEQUAL SIZES When the cluster is some existing human group, it will usually contain varying numbers of elements. This will be true of dwellings in blocks, employees in work groups, or students in classes. There are several important consequences. The planning and administration of the study must reckon with the fact that it has no exact control over the size (n) of the sample. For example, the sample of 25 blocks, mentioned earlier in this section, will generally not contain exactly the planned 400 dwellings but will contain more or fewer, depending on the sample of blocks that happened to be selected. Sometimes the variation in the size of the sample is reduced through the use of information on the sizes of the individual clusters (see Section 20). Let us assume, however, that we are talking about a simpler model in which m clusters are taken with random choice out of the M clusters which compose the population. The estimate of the mean is usually calculated as the simple mean of the sample cases $\bar{x}_c' = \sum x/n$. The estimate of the mean may be seriously biased if m is small and there is inequality in the size of the clusters (12, Chap. 6). Thus, the sample mean is no longer equal to a simple mean of the cluster means, as it was for equal sized clusters. However, a useful analogy may be found in that the sample mean is equal to a *weighted* mean of the cluster means, the weights are the relative sizes of the clusters. Thus

$\bar{x}_c' = \frac{1}{m} \sum \frac{N_i}{\bar{N}'} \bar{x}_i$. Here $\bar{N}' = \frac{n}{m} = \frac{1}{m} \sum N_i$, the average number of elements in the sample clusters. Hence N_i/\bar{N}' is the size of the cluster relative to the average size in the sample.

The variance of this estimate has an analogous form to the variance shown before.

$$\text{est var } (\bar{x}_c') = (1 - f) \frac{1}{m} \left[\frac{1}{m-1} \sum \left(\frac{N_i}{N'} \right)^2 (\bar{x}_i - \bar{x}_c')^2 \right]$$

That is the squared deviations now are multiplied by the squares of the relative weights of the clusters

13 *The Effects of Clustering, Intraclass Correlation*

Of the various modifications of sample design (listed in Section 7), clustering usually has by far the greatest effect. If we compare a sample of n independently selected elements with another sample containing a like number of elements but selected in m clusters, we note that the number of independent choices involved is reduced considerably. Although the former sample would be well spread over the population, the latter would be bunched in spots. In the example in Section 12, the sample of dwellings was confined to only 25 of the city's 750 blocks. This clustering would be of no consequence if all the elements in the population were scattered at random into the different clusters. In most practical cases, however, we find that the elements in a cluster tend to be more like other elements in the same cluster than like elements in other clusters. The various dwellings in the same block will show a greater homogeneity than a similar number of dwellings scattered throughout the city. The measure of this homogeneity is the intraclass correlation.

In the example of newspaper subscribers of Section 12, we found that the estimate \bar{x}_{cc}' based on 40 clusters of 10 subscribers each was subject to a variance which was estimated as

$$\text{est var } (\bar{x}_{cc}') = (1 - f) \frac{s_b^2}{m} = .001217$$

The standard error was estimated at $\sqrt{.001217} = .035$

Suppose that after this cluster sample was selected the researcher mistakenly used the formula for the standard error of a simple random sample, which is inappropriate for the sample design. What is the consequence of this mistake? In our example he would take (see Section 1)

$$\sqrt{(1-f) \frac{p'(1-p')}{n-1}} = \sqrt{.967 \times \frac{.37 \times .63}{399}} = \sqrt{.000565} = .024;$$

$$\text{also } 1.96 \times .024 = .047.$$

Hence he makes the statement that the population value was within the limits of 37 percent \pm 4.7 percent, and he would say that this statement has but a 5-percent chance of being incorrect. We are wiser; we know that the true standard error is .035. Hence 4.7 percent is equal not to 1.96 standard errors but only to $.047/.035 = 1.34$ standard errors. In the appropriate tables we find that the probability of a statement based on a confidence interval of 1.34 times the standard error has a probability of 0.18 of being incorrect. Hence, the consequence of the researcher's mistake is that his confidence, and that of his readers, in his results is misplaced: his statements will be incorrect in the long run not 5 times in 100 but 18 times in 100.

The effect of clustering on sample results is often serious. Yet it is disregarded very frequently. It is quite common to see reports of estimates which grossly underestimate the actual variability of the sample and consequently grossly exaggerate the significance of the sample results. Marks (17) found that for the revision of the Stanford-Binet scale the ratio of the correct standard error of the mean I.Q. to the incorrect s/\sqrt{n} is 3.36. Therefore, the researcher using 1.96 of his incorrect "standard errors" is using but $1.96/3.36 = 0.58$ of the actual standard error. With the use of the incorrect intervals the researcher would be making incorrect statements not 5 times in 100, as he hoped, but 56 times in 100.

A convenient measure of the effect of clustering can be given in terms of the coefficient of intraclass correlation, ρ . This actually is based on the ratio of two variances, and it can be estimated from two values we already have. The variance of our cluster sample is: $(1-f) s_b^2/m = .001217$. The value of $(1-f) s^2/n = .000565$ is a usable estimate for the value of the variance to which a simple random sample of the same population would be subject. Their ratio is:

$$\frac{s_b^2/m}{s^2/n} = 1 + \text{est. } \rho (A - 1).$$

In our example, A (the number of elements in each cluster) is 10. Hence we have

$$1 + \text{est } \rho (10 - 1) = \frac{001217}{000565} = 2.15,$$

$$\text{and est } \rho = \frac{2.15 - 1}{10 - 1} = +.13$$

That is, the variance of the cluster sample in this case is estimated as 2.15 times greater than the variance one would expect from a simple random sample of the same number of elements. This ratio of the variances can be expressed as due to an estimated ρ of +.13.

When ρ is positive, the ratio of the variances is greater than 1 and the cluster sample has a greater variance than a simple random sample of the same size would have. The maximum possible value for ρ is +1, in which case the ratio has a value of A . This corresponds to a case of complete segregation of a characteristic: all the individuals of every cluster are exactly alike with regard to that characteristic. Note, also, that the effect of clustering is equal to $\rho(A - 1)$, hence, a relatively small ρ may have a serious effect on the variance if the size of the cluster (A) is large. In the case of clusters of unequal sizes, the average size \bar{N}' of the cluster can be used in place of A . In a subsampling design (Section 19) the average number of subsampled elements per cluster plays the same role.

The variance of a cluster sample is almost always greater than that of a simple random sample of the same size. However, this is a sociological fact and not a logical necessity. For example, if balls are selected from a "well mixed urn," it makes no difference whether one takes them in clusters or singly, the balls will be randomly distributed in the clusters. In this situation we would expect a value of 1 for the ratio of the variances and a value of zero for ρ . However, people do not enter human groups "well mixed." Furthermore, although a negative ρ is possible, it is a rare phenomenon for social variables. The lowest possible value is $-\frac{1}{A - 1}$, corresponding to a zero value for the ratio.

For most human groups, ρ tends to be positive. That is, the individuals associated with human groups tend to resemble one

another. These tendencies to homogeneity in groups hold for most characteristics. The homogeneity of people—the segregation of characteristics—is greater than would exist if people were selected into groups by random choice. These tendencies may be due to selection, to mutual influence, to the correlation among characteristics, or to a combination of these. In any case, the tendency is a social phenomenon in that ρ is a measure that belongs to the group as such, it has no meaning for the individual except in so far as he is considered as a member of a group (or cluster). ρ is a measure of the fraction of the total variance which elements in the same sampling unit, the same cluster, have in common. It may be looked on as measuring the amount of homogeneity or segregation of the elements within a group of units. Given a set of elements, the greater the segregation into separate units of similar elements, the greater is the homogeneity within the unit. It should be of great utility in the social sciences as a measure for description and comparison. Of course, ρ is specific for a characteristic. For the same grouping of individuals, different variables will exhibit different ρ s. Also, for any specific characteristic, the value of ρ depends on the actual distribution of the population into the groupings being considered.

That individuals are segregated in social groups is of importance to the sampler because he often uses those groups (or some approximation thereof) as sampling units in the process of selection. Therefore, the variance of sample results will be influenced by the clusters used in the selection process. We see, then, that the precision of the results of a cluster sample cannot be given simply in terms of the numbers of elements (cases) in the sample. The numbers of the other units selected—*i.e.*, the numbers of the different kinds of clusters used—will be of importance also.

PRACTICAL PROCEDURES

14 Economy

The results of a sample based on clusters are almost always subject to greater sampling error than those of a sample of the same number of elements selected individually. Why, then, use clustering in selection?

Because the cost per element in cluster sampling is less, often substantially so, than for a sample of individually selected elements Clustering should be preferred over individual selection in so far as the lowering of the cost per element due to clustering is greater than the increase in the variance per element

Let us begin with the data of the 400 subscribers in 40 sample clusters in Sections 12 and 13 Suppose that the same characteristic is to be measured again next year and the cost of the study is specified We want to compare two sample designs in order to use the one which yields the smaller variance for the available expenditure First, estimate how many sampling units can be obtained for the available money under each of the two designs Suppose that it is estimated that with one design 300 subscribers in 30 clusters may be obtained, or one could get 100 subscribers selected individually Which of those two samples will have the smaller variance? The variance of the cluster sample is

$$\text{approx var } (\bar{x}_{ec}') = \frac{s_b^2}{m}$$

The factor of $(1 - m/M)$ is neglected in this discussion, and it is seldom of any consequence (see Section 1) We found in Section 12 that in our example the variance of individual clusters is estimated as $s_b^2 = 0504$ Therefore, the variance of a sample of 30 clusters is estimated as $0504/30 = 0017$

The variance of the simple random sample is

$$\text{approx var } (x_o') = \frac{s^2}{n}$$

Again, we neglect the factor $(1 - f)$ Since we deal with a proportion

$$s^2 = \frac{np'(1-p')}{n-1} = \frac{400(37)(63)}{399} = 234$$

This is a usable estimate of the variance of individual elements (The formula was given in Section 1, and the value of 37 was used in Sections 12 and 13) Therefore, the variance of a sample of 100 elements is estimated as $\frac{234}{100} = 00234$ Since for the available money the cluster sample yields the smaller variance, the greater precision, it should be preferred

Alternately, one may begin with a fixed precision that he wants the sample to have. Suppose, now, that our purpose is again to measure the same characteristic on another occasion but now it is specified that a precision of 6 percent is required at the 0.95 probability level. This means that a standard error of 3 percent = 0.3 is required. That is, the variance of the sample mean is to be $(0.3)^2 = 0.09$. Which of the two designs will yield a sample of that precision for less cost? Using the equation for the variance of a cluster sample, we find the number of clusters necessary for the desired variance of 0.09

$$0.09 = \frac{0.504}{m} \text{ or}$$

$$m \text{ for variance of } 0.09 = \frac{0.504}{0.09} = 56 \text{ clusters}$$

Using the equation for the variance of a simple random sample of elements, we find similarly that n for variance of 0.09 = $234 / 0.09 = 260$ elements †

Now the cost of visiting 560 dwellings in 56 clusters is to be estimated, also, the cost of 260 dwellings selected individually. Of these two designs, either of which would yield the same variance, we choose that which costs less. Note that the required precision dictates a sample size, in numbers of elements or clusters, and not a sampling rate. For a desired precision and for a given variability among the sampling units of the population, the required sample size is about the same for a small as for a large population.

Two views of economy have been presented for a fixed variance we sought the less costly of two designs, then for a fixed cost we sought the design with the smaller variance. In both cases we sought the design which yields the least variance (i.e., the most precision, the most information) per unit cost. The superlative in the last sentence is justified because the search for economy is not limited usually to two alternatives only, the principles and procedures discussed hold for

†It takes $\frac{560}{260} = 2.15$ as many elements in the cluster sample as in the sample of individuals to obtain the same variance. This is the same ratio as we found in Section 13 for the ratio of the variances of the two designs with equal numbers of elements. That fraction is determined by the size (A) of the cluster and the intraclass correlation (ρ) for the characteristic in the specific clustering.

the comparison of any number of alternative designs. For example, in the case above we might investigate whether a cluster of 3 or 5 or 20 elements might be superior to both single elements and to clusters of 10.

If the cost per element were the same for the clusters of 10 elements as for the individual elements, the latter would usually show up better. It will generally yield the same variance for smaller n or the smaller variance for the same n . However, the cost per element is generally not the same for elements in clusters as for elements selected individually. Hence, the comparison should be made between variances per unit cost.

As these comparisons are made, it becomes obvious that the number of elements—say, interviews—in the sample is not the sole important factor as regards either the cost or the variance of sample designs. The numbers of all other sampling units selected—such as dwellings, blocks, towns, cities, counties—are also relevant. Furthermore, other aspects of design—stratification, varying probabilities of selection, and methods of estimation—should be considered also. However, in most social studies those aspects are not so important as the effect of clustering, either for the variance or for the costs of the study.

Three important questions for which we do not have definite answers are pertinent:

(a) Where may we get the needed estimates of variances and of cost factors? We can prepare them on the basis of past experience with similar surveys, ask an expert or conduct a pilot study. Good estimates of cost factors are especially difficult to get.

(b) How do we find the precision needed? The user of the statistic should determine this. Usually this factor is a much greater source of uncertainty than those in (a). Theoretically, one might say that a sample is too large if the statistics it yields are more precise than is warranted by the uses of those statistics, that a sample is too small if its statistics are not precise enough to aid in making decisions based on them to an extent commensurate with its cost. According to this rational picture, the desired precision would always determine the sample size, and the necessary cost would be allocated in accord with that determination. Actually, we seldom find the necessary precision so well defined—or the available funds so fluid.

(c) For studies with several or many important characteristics to measure, how do we evaluate the relative importance of each? How

do we arbitrate between the conflicting answers on the desirable size and design of the sample? This area of decisions is even more obscure than the preceding one

Despite these difficulties, theory does provide general outlines for the decision on the required sample size and it can be of great help in choosing an efficient design. Fortunately, moderate departures from optimum design do not incur heavy losses in economy. Hence, rough guesses may serve in place of better estimates, and useful compromises between conflicting aims in design can often be made.

The designing of samples is mainly an engineering job, the available theory and the knowledge of results with similar materials should be utilized to produce a desired result with the available resources and with the greatest economy. This, at any rate, should be the goal, but the use of the superlative in the preceding sentence is immodest—it usually represents a level of aspiration rather than of achievement. In making economy the aim, we should understand that cost is to be understood broadly to mean effort in general. Since effort as well as money available for research is limited, economy helps to increase the total quantity and quality of research output (1, p. 50ff).

15 *Practicality*

A probability sample cannot be created by assumption, nor will it be "given," as in the examples of elementary statistics. The dictum of "quota" samplers to their interviewers, "Go out and get a random sample," is most impractical. The interviewer is not capable of doing it, nor is his dispatcher. The need for a mechanical method of selection has been stated in Sections 4 and 5. Now we want to emphasize the need for taking care to translate the theoretical model of selection into a complete set of simple, practical instructions.

It is necessary to give the field interviewer simple and clear instructions for the carrying out of his tasks. The less attention the sampling instructions demand of him, the more he can devote to his principal and difficult task of interviewing. For example, in order to identify a sample segment the interviewer should not be asked to locate a long arbitrary straight line marked on a map—but he can locate a street. His sampling instructions should be confined to locating streets and addresses, listing occupants of the household, and so on. These tasks are difficult enough in some cases.

Of course, "clear, simple, practical and *complete* instructions" represent aims rather than fulfillment. Actually, in the fitting of practical field sampling procedures of large-scale samples to the statistical model of the design, there will often be some unfilled gaps (see Section 16). Part of the art of practical work is in guessing what irregularities, where and how much, one can afford to tolerate.

The sample design is no better than the weakest link in the entire procedure. Each sample design is an adaptation of sampling theory to the resources at hand. The resources include the distribution of the population, the facilities for communication, the nature and training of the field force, and the researchers engaged in the task. They also include the receptivity of the administration as well as of the users of research—their receptivity to, and understanding of, the methodological tools.

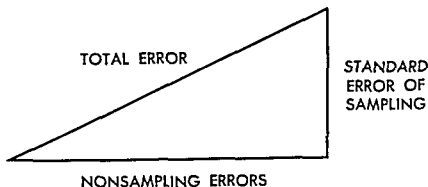
16 *Nonsampling Errors*

In this chapter the verbal statements of the confidence interval refer to "the population value." This is defined as the value that would have been obtained if the entire population—rather than just a sample—had been designated for observation. This definition deliberately avoids the "true value," the parameter, because there are sources of error which exist even if every element is designated for observation. These errors are called nonsampling errors, to distinguish them from the sampling errors which arise because only a part of the total population is designated for observation. The nonsampling errors are sometimes called errors of observation, or errors of measurement, or errors of response (3, pp 15-52, 22, pp 9-16). They occur because observations have to be made to obtain some needed result and because the physical procedures of observations are subject to imperfections. The sampling errors occur when the observations are made on only a fraction of the population.

Nonsampling errors may be of two types: *variable-response errors* and *biases*. Within the variable response error are included all those errors in the procedures of observation (interviewing, coding, punching, nonresponse, etc.) which tend to cancel each other in the long run. On the contrary, we include under the term *bias* all the discrepancies between our observations and the quantities we aim to measure of the systematic noncanceling type. The size of this bias is unknown in

practice. Thus the population value obtained in a census by means of a single observation on each individual is subject to error. Because of the theoretical fluctuation due to the variable-response error, we may think of a distribution of possible population values. The mean of this theoretical distribution of population values is the "expected population value." The difference between this value and the "true value" of the parameter is the bias (1, pp 292-317, 11, pp 147-154).

In addition to the nonsampling errors of bias and of variable response, sample studies are also subject to sampling errors. In the general study design, all of these errors may be considered together as constituting the *total error of the sample*. This total error is the square root of the sums of squares of two quantities. The first is the standard error of the sampling distribution, the *sampling error*. The second quantity is the combined effect of the two kinds of nonsampling errors which we called variable-response error and bias. That is $(\text{total error})^2 = (\text{standard error of sampling})^2 + (\text{nonsampling errors})^2$. This relationship may be illustrated by means of the three sides of a right triangle.



The total error depends on the length of both of the legs and cannot be shorter than either of them. The standard error leg can be shortened sometimes by a change in sample design, and always by taking more sampling units—either more clusters or more individuals. Of the nonsampling errors, the variable-response errors may be reduced either by taking more of something—observations per individual, or individuals, or interviewers—or by improving the precision of the methods of observation. But the length of the nonsampling leg may be due mostly to bias, which can be reduced only through better survey procedures.

through improving the questionnaire, or the field work, or the coding and processing, etc.¹

It may be wasteful to spend much money on a large sample in order to reduce the standard error if the nonsampling errors are allowed to remain large and vice versa. In the general study design, the nonsampling errors should be considered together with the sampling error, because together they constitute the total error of the survey.

An important special class of nonsampling errors in social studies is composed of the errors of nonresponse. These arise whenever a member of the population designated for the sample is not included in the results—because his answer was “not ascertained,” because of total refusal of the interview, because of not being at home, because of illness, or for similar reasons. This is a nonsampling error, it can occur even if the entire population is designated initially for the study. In order to reduce the nonresponses, most surveys must make several call backs (1, pp 292-304). The reduction of nonresponse is an important aim in many social studies. The wording and the structure of the interview and of the schedule is planned with the aim of reducing the incidence of refusals. The time of calls in homes is planned so as to reduce the incidence of not at homes.

Our statements of statistical inference were made about the population values. In addition to the stated estimates of the sampling error, we must consider sample results as subject to unstated nonsampling errors.

The measurement of sampling errors is a matter of the proper planning of the sample design, and of the proper calculation of sample estimates. But the measurement of nonsampling errors requires special procedures, and it is usually a costly addition to the main objectives of the survey. The reduction of nonsampling errors is a constant challenge to the researcher which calls for long range research.

¹As the result of two dichotomies, we may classify the errors of the survey as (a) sampling errors (b) sampling biases (c) nonsampling variable errors of response, and (d) nonsampling biases. It is understood that (a) is in one leg and (c) and (d) in the other. What about (b), the sampling biases? It will be convenient to consider biases in the selection procedures included with other procedural biases in our definition of bias. Different from these are any biases due to the procedures of estimation (see Section 22), in most practical cases they are either nonexistent, or comparatively slight if the sample is not very small (1, pp 111-159). Finally, it should be noted that the model above represents but one of the many possible ways for classifying the great number of possible sources of error to which sample results may be subject. In other places the reader will find other useful classifications (1 pp 292 317, 3, p 129).

VARIOUS STATISTICS

17 *About Various Statistics*

The discussions in this chapter were built around problems of estimating the mean of a population. The restriction was for the sake of simplicity and convenience. The choice of the mean rests on its basic importance. The estimation of totals for the population is closely related to that of means (as discussed in Section 1). Furthermore, proportions are but a special case of the mean, and a great deal of social research is reported in terms of proportions, of frequency distributions in classes. The distributions may be of attributes, behaviors, attitudes, or opinions, they very often are expressed as proportions of the total.

In addition to the mean for the entire population total, the estimation of means for subdivisions of the total population is often of great importance. These have been called "domains." "Any subdivision about which the enquiry is planned to supply numerical information of known precision may be termed a domain of study" (21, p. 5). In general, the principles discussed in this chapter in terms of means of the total population apply as well to the means of the domains. Hence, the results based on cross-tabulations present no new problems in principle. However, it is true that, compared with the total population, in dealing with a domain one gets more often into "cells" so small that the problems of small samples, particularly the questions of non-normality, become important. Hence, the reference to available tests for "distribution-free" estimates is particularly relevant here (see Chap. 12). Furthermore, the effect of clustering becomes less drastic for many of the domains than for the entire sample because of the smaller number of elements per cluster.

Sometimes a researcher will say that the question of sampling is of no interest to him because he wants not to estimate quantities but merely to measure relations. This view may overlook the fact that the relationships are measured in terms of statistics—in comparisons of proportions, in correlation coefficients, etc. These statistics, too, depend on the individuals included in the sample. When some relationship is expressed in terms of a number based on sample data, that number is a statistic, a sample estimate of a population characteristic. The statistic is subject to sampling error, and the sampling error can be expressed in terms of a confidence interval.

through improving the questionnaire, or the field work, or the coding and processing, etc.¹

It may be wasteful to spend much money on a large sample in order to reduce the standard error if the nonsampling errors are allowed to remain large, and vice-versa. In the general study design, the nonsampling errors should be considered together with the sampling error, because together they constitute the total error of the survey.

An important special class of nonsampling errors in social studies is composed of the errors of nonresponse. These arise whenever a member of the population designated for the sample is not included in the results—because his answer was “not ascertained,” because of total refusal of the interview, because of not being at home, because of illness, or for similar reasons. This is a nonsampling error, it can occur even if the entire population is designated initially for the study. In order to reduce the nonresponses, most surveys must make several call backs (1, pp 292-304). The reduction of nonresponse is an important aim in many social studies. The wording and the structure of the interview and of the schedule is planned with the aim of reducing the incidence of refusals. The time of calls in homes is planned so as to reduce the incidence of not at homes.

Our statements of statistical inference were made about the population values. In addition to the stated estimates of the sampling error, we must consider sample results as subject to unstated nonsampling errors.

The measurement of sampling errors is a matter of the proper planning of the sample design, and of the proper calculation of sample estimates. But the measurement of nonsampling errors requires special procedures and it is usually a costly addition to the main objectives of the survey. The reduction of nonsampling errors is a constant challenge to the researcher which calls for long range research.

¹As the result of two dichotomies, we may classify the errors of the survey as (a) sampling errors (b) sampling biases (c) nonsampling variable errors of response, and (d) nonsampling biases. It is understood that (a) is in one leg and (c) and (d) in the other. What about (b), the sampling biases? It will be convenient to consider biases in the selection procedures included with other procedural biases in our definition of bias. Different from these are any biases due to the procedures of estimation (see Section 22), in most practical cases they are either nonexistent, or comparatively slight if the sample is not very small (1, pp 111-159). Finally, it should be noted that the model above represents but one of the many possible ways for classifying the great number of possible sources of error to which sample results may be subject. In other places the reader will find other useful classifications (1 pp 292-317, 3, p 129).

VARIOUS STATISTICS

17 *About Various Statistics*

The discussions in this chapter were built around problems of estimating the mean of a population. The restriction was for the sake of simplicity and convenience. The choice of the mean rests on its basic importance. The estimation of totals for the population is closely related to that of means (as discussed in Section 1). Furthermore, proportions are but a special case of the mean, and a great deal of social research is reported in terms of proportions, of frequency distributions in classes. The distributions may be of attributes, behaviors, attitudes, or opinions, they very often are expressed as proportions of the total.

In addition to the mean for the entire population total, the estimation of means for subdivisions of the total population is often of great importance. These have been called "domains." "Any subdivision about which the enquiry is planned to supply numerical information of known precision may be termed a domain of study" (21, p. 5). In general, the principles discussed in this chapter in terms of means of the total population apply as well to the means of the domains. Hence, the results based on cross-tabulations present no new problems in principle. However, it is true that, compared with the total population, in dealing with a domain one gets more often into "cells" so small that the problems of small samples, particularly the questions of non-normality, become important. Hence, the reference to available tests for "distribution-free" estimates is particularly relevant here (see Chap. 12). Furthermore, the effect of clustering becomes less drastic for many of the domains than for the entire sample because of the smaller number of elements per cluster.

Sometimes a researcher will say that the question of sampling is of no interest to him because he wants not to estimate quantities but merely to measure relations. This view may overlook the fact that the relationships are measured in terms of statistics in comparisons of proportions, in correlation coefficients, etc. These statistics, too, depend on the individuals included in the sample. When some relationship is expressed in terms of a number based on sample data, that number is a statistic, a sample estimate of a population characteristic. The statistic is subject to sampling error, and the sampling error can be expressed in terms of a confidence interval.

"Tests of significance" of sample results are basic in research. The statistical tools required for many tests of significance are the same as those which are necessary for the construction of the analogous confidence intervals.* Suppose, for example, that it is desired to make a test of significance between two means, each pertaining to a domain of the study. The statistic used in the test is a ratio, the numerator is the difference of the two means, and the denominator is the standard error of that difference. The variance of the difference for the case of two independent simple random samples is simply the sum of the variances of the two sample means. For more complex designs, another term has to be considered—the covariance of the two sample means brought about by the design.

A frequently used test of the presence of relationship is the chi-square test based on the 2×2 cells of two dichotomies. In the case of simple random samples, that test is very similar to the test of the difference of two proportions (20, p. 203). In case of a more complex design, the chi-square test is not valid because its underlying assumptions of the independent selection of sample cases is violated by the sample design. But the test of the difference of the two proportions may be made. This requires that the correlations due to clusterings and to the other complexities of the design be considered.

COMPLEX SAMPLE DESIGNS

18 *Stratified Cluster Sampling*

In Section 12 we dealt with a sample of subscribers as an example of cluster sampling. The estimated variance of the sample mean was given as $(1 - f) s_b^2/m$. That formula is applicable to the case where the sample of m out of M clusters represents m independent random selections. It is basic to the more complex designs. However, the simple random choice of clusters is not used frequently in practice. Stratification is used generally. Very frequently the selection is systematic, as in our example of the clusters of subscribers. If the listing of the clusters had been random—as after a thorough shuffling—the

*Because of the logical basis of the null hypothesis which underlies tests of significance, the correction for finite population $(1 - f)$ should not be included in the variance formulas when these are used in tests of significance (3, p. 247).

systematic selection would be equivalent to a random choice. It was stated specifically, however, that neighboring routes were filed next to one another. Therefore, the similarities, slight or great, that might exist in neighboring routes would be reflected in a selection of clusters rather evenly spread over the different neighborhoods of the area. We shall assume that the chief effect of this systematic selection of the clusters is to yield a stratified sample of them (see Section 10).

Each selection came from its own "implicit stratum" of 30 clusters (the interval was 1 in 30). However, we shall use an approximation called the method of "collapsed strata" and assume that each successive set of four selections was selected by random choice from a group of 120 clusters. Thus, we shall have 10 strata with the equal weight of $1/10$ for each. There are four equal-sized clusters selected from each stratum. The sample mean is calculated as before, but the sample design now assumed calls for a different formula for obtaining the estimated variance of the sample mean. We need here a specific application of the general formula for stratified samples (Section 8)

$$\text{est var } (\bar{x}_w') = \sum w_h^2 [\text{est var } (\bar{x}_h')]$$

The quantity within the brackets is the variance within one of the strata. Within each stratum we have a sample of four equal clusters out of 120. Hence (as in Section 12) we have in the h th stratum

$$\text{est var } (\bar{x}_{ec}')_h = \left[\left(1 - \frac{1}{30} \right) \frac{1}{4} \frac{1}{3} \sum (\bar{x}_i - \bar{x}_{ec}')^2 \right]_h$$

Then, for the entire sample, we have

$$\begin{aligned} \text{est var } (\bar{x}_{ecw}') &= \sum \left(\frac{1}{10} \right)^2 [\text{est var } (\bar{x}_{ec}')_h] \\ &= \sum \left(\frac{1}{10} \right)^2 \left[\left(1 - \frac{1}{30} \right) \frac{1}{4} \frac{1}{3} \sum (\bar{x}_i - \bar{x}_{ec}')^2 \right]_h \\ &= \frac{29}{30} \frac{1}{100} \frac{1}{4} \frac{1}{3} \sum \sum (\bar{x}_i - \bar{x}_{ec}')^2 \end{aligned}$$

In the above, $\sum \sum (\bar{x}_i - \bar{x}_{ec}')^2$ is the term we need to calculate

"Tests of significance" of sample results are basic in research. The statistical tools required for many tests of significance are the same as those which are necessary for the construction of the analogous confidence intervals.* Suppose, for example, that it is desired to make a test of significance between two means, each pertaining to a domain of the study. The statistic used in the test is a ratio, the numerator is the difference of the two means, and the denominator is the standard error of that difference. The variance of the difference for the case of two independent simple random samples is simply the sum of the variances of the two sample means. For more complex designs, another term has to be considered—the covariance of the two sample means brought about by the design.

A frequently used test of the presence of relationship is the chi-square test based on the 2×2 cells of two dichotomies. In the case of simple random samples, that test is very similar to the test of the difference of two proportions (20, p. 203). In case of a more complex design, the chi square test is not valid because its underlying assumptions of the independent selection of sample cases is violated by the sample design. But the test of the difference of the two proportions may be made. This requires that the correlations due to clusterings and to the other complexities of the design be considered.

COMPLEX SAMPLE DESIGNS

18 *Stratified Cluster Sampling*

In Section 12 we dealt with a sample of subscribers as an example of cluster sampling. The estimated variance of the sample mean was given as $(1 - f) s_b^2/m$. That formula is applicable to the case where the sample of m out of M clusters represents m independent random selections. It is basic to the more complex designs. However, the simple random choice of clusters is not used frequently in practice. Stratification is used generally. Very frequently the selection is systematic, as in our example of the clusters of subscribers. If the listing of the clusters had been random—as after a thorough shuffling—the

*Because of the logical basis of the null hypothesis which underlies tests of significance, the correction for finite population $(1 - f)$ should not be included in the variance formulas when these are used in tests of significance (3, p. 247).

systematic selection would be equivalent to a random choice. It was stated specifically, however, that neighboring routes were filed next to one another. Therefore, the similarities, slight or great, that might exist in neighboring routes would be reflected in a selection of clusters rather evenly spread over the different neighborhoods of the area. We shall assume that the chief effect of this systematic selection of the clusters is to yield a stratified sample of them (see Section 10).

Each selection came from its own "implicit stratum" of 30 clusters (the interval was 1 in 30). However, we shall use an approximation called the method of "collapsed strata" and assume that each successive set of four selections was selected by random choice from a group of 120 clusters. Thus, we shall have 10 strata with the equal weight of $1/10$ for each. There are four equal sized clusters selected from each stratum. The sample mean is calculated as before, but the sample design now assumed calls for a different formula for obtaining the estimated variance of the sample mean. We need here a specific application of the general formula for stratified samples (Section 8)

$$\text{est var } (\bar{x}_w') = \sum w_h^2 [\text{est var } (\bar{x}_h')]$$

The quantity within the brackets is the variance within one of the strata. Within each stratum we have a sample of four equal clusters out of 120. Hence (as in Section 12) we have in the h th stratum

$$\text{est var } (\bar{x}_{ec}')_h = \left[\left(1 - \frac{1}{30} \right) \frac{1}{4} \frac{1}{3} \sum (x_i - \bar{x}_{ec}')^2 \right]_h$$

Then, for the entire sample, we have

$$\begin{aligned} \text{est var } (\bar{x}_{ecw}') &= \sum \left(\frac{1}{10} \right)^2 [\text{est var } (\bar{x}_{ec}')_h] \\ &= \sum \left(\frac{1}{10} \right)^2 \left[\left(1 - \frac{1}{30} \right) \frac{1}{4} \frac{1}{3} \sum (x_i - \bar{x}_{ec}')^2 \right]_h \\ &= \frac{29}{30} \frac{1}{100} \frac{1}{4} \frac{1}{3} \sum \left[\sum (x_i - \bar{x}_{ec}')^2 \right]_h \end{aligned}$$

In the above, $\sum \sum (x_i - \bar{x}_{ec}')^2$ is the term we need to calculate

There are ten terms, of which the first is

$$[\sum (\bar{x}_i - \bar{x}_{ec})^2]_1 = (4 - 325)^2 + (7 - 325)^2 + (1 - 325)^2 + (1 - 325)^2 = 2475$$

$$\text{Here } (\bar{x}_{ec})_1 = \frac{1}{4} (4 + 7 + 1 + 1) = \frac{1}{4} (13) = 325$$

The sum of the ten terms (of which the first is 2475) equals 1 505. Hence

$$\text{est var } (\bar{x}_{ec} - w) = 967 \frac{1}{1200} 1 505 = 00121$$

In our case this approximation did not produce a variance smaller than the 001217 obtained without considering any strata in Section 12.

We might have used 20 strata of two clusters each. Then the variance would come from the 20 differences between pairs of clusters. A formula used in systematic sampling, and mentioned in Section 10, utilizes all of the 39 successive differences. That is

$$\begin{aligned} \text{est var } (\bar{x}_{ec} - w) &= \left(1 - \frac{1}{30}\right) \left[\frac{1}{40} \frac{1}{2} \frac{1}{39} \sum_{39} (\bar{x}_i - \bar{x}_{i+1})^2 \right] \\ &= \frac{29}{30} \frac{1}{40} \frac{1}{2} \frac{1}{39} \frac{1}{100} \sum_{39} (X_i - X_{i+1})^2 = \\ &= 967 \frac{1}{312000} 344 = 00107 \end{aligned}$$

$$\text{Here } \sum_{39} (X_i - X_{i+1})^2 = (4 - 7)^2 + (7 - 1)^2 + (1 - 1)^2 + (5 - 2)^2 + (2 - 4)^2 = 344$$

The three separate calculations of the variance of this sample did not yield appreciably different results.

†It is easier to calculate the equivalent term

$$\frac{1}{100} \left[\sum X_i^2 - \frac{1}{4} (\sum X_i)^2 \right] = \frac{1}{100} \left[(16 + 49 + 1 + 1) - \frac{1}{4} (13)^2 \right] = 2475$$

The sum of the ten terms is

$$\begin{aligned} \frac{1}{10} \left\{ \frac{1}{100} \left[\sum X_i^2 - \frac{1}{4} (\sum X_i)^2 \right] \right\} &= \frac{1}{100} \left[\sum X_i^2 - \frac{1}{4} \sum (\sum X_i)^2 \right] \\ &= \frac{1}{100} \left[744 - \frac{1}{4} 2374 \right] = 1 505 \end{aligned}$$

In Sections 19, 20, and 21, the selection of subsamples is discussed. The examples discussed in each of these sections deal with samples to which the calculation of the simple mean $\sum x/n$ is proper. The calculation of variances for subsamples in general will not be discussed in this chapter. However, if the subsamples in all the clusters are of equal size (or very nearly so), the formulas for the variances of the mean are similar to those given just above. Equal-size subsamples are obtained with systematic selection with probabilities proportional to size. An example from Section 20 is the subsample of 10 employees each from a sample of 40 work groups. A satisfactory approximation might be:

$$\begin{aligned}\text{est. var. } (\bar{x}_{e..})' &= \frac{1}{2m(m-1)} \sum^{m-1} (\bar{x}_i' - \bar{x}_{i+1}')^2 \\ &= \frac{1}{2 \times 40 \times 39} \sum^{39} (\bar{x}_i' - \bar{x}_{i+1}')^2.\end{aligned}$$

The use of \bar{x}_i' rather than \bar{x}_i denotes the fact that, since we have only a sample of the employees of any section, we deal not with the exact value of \bar{x}_i but only with its estimate \bar{x}_i' . The subscript $e..$ denotes a sample composed of subsamples of equal size.

No formulas for the variance will be given in this chapter for stratified samples based on clusters or on cluster subsamples which are of different sizes. The possibilities for varieties of procedures both of selection and of estimation are very great. Differences which may appear minor to the unpracticed eye may have important effects on the design. It is wise to consult a sampling statistician with the design during the planning stage (1, pp. 234-267).

19. *Subsampling. A Two-stage Sample*

Suppose that a sample is to be taken of a city's dwellings and that two designs are compared, with the aim of making a choice between them. One calls for 400 dwellings to be selected individually from the listing of the city's total of 12,000; the second calls for a cluster sample of 25 of the city's 750 blocks (as described in Section 12). In each case we have a sampling fraction of 1 in 30, a probability of 1 in 30 for every dwelling in the city. We are comparing two samples with the same expected number of elements. The objection to the sample of

individual dwellings is that its cost per element is high. There are an average of 16 dwellings per block, hence it takes on the average almost two blocks to produce a sample dwelling. The cost of locating the dwellings may not be serious within this moderate-sized city, although in other cases of clustering the cost of travel may be of the greatest importance (see Section 21). However, no satisfactory listing of the dwellings exists. Thus 29 nonsample dwellings would need to be listed for every sample dwelling and this additional burden increases the cost per element appreciably.

The chief objection to the cluster design is that the variance per element is high. An average of 16 dwellings per block means that the ratio of the variance of this sample to the first design is $[1 + \rho(16 - 1)]$ (see Section 13). Hence, for example, if for a characteristic ρ is 0.1, then the variance of the cluster design will be 2.5 times greater than that of the first design. The trouble is that with the cluster design the sample is confined to only 25 blocks in the city. The spread of the sample is very restricted.

This is the general rule: with the increase in the size of the cluster the cost per element decreases but the variances per element of the sample estimates increase (see Section 14). We may then look for a compromise between these two factors: on the one hand, we wish to spread our sample into as many clusters as possible in order to include in the sample the diverse elements in the population, on the other, we wish to keep down the costs incurred by covering many clusters. In other words, we wish to look for that cluster size which yields a good compromise between the two conflicting effects of clustering. That is, after comparing[†] the economy of the two designs, we look at other designs—in this case for some compromise in the form of smaller clusters. However, the sizes of clusters are often such that it is inconvenient to split them. For instance, most city blocks have streets for boundaries, so that their areas can be located with relatively simple instructions, but a fraction of a block is hard to define.[‡]

Suppose that instead of the sample of entire blocks we investigate

[†]Comparison is in the manner discussed in Section 14. The present discussion is in terms of a constant number of elements (n), but it may be translated into the more rational grounds of a fixed allowable variance or cost.

[‡]In Section 21 another reason for subsampling appears, an increase in the effective spread of the sample without incurring a proportionate increase in costs (3, p. 189).

a third design that would triple the number of clusters in the sample it would take $3 \times 25 = 75$ blocks, and an average of $16/3 = 5.3$ dwellings in each. The intraclass correlation does not change, since it is a property of the actual block but its effect becomes less than one third as great as before. In a calculation of the effect of clustering, the size of the average subsample taken from the cluster appears as the multiplier. For the case of $\rho = 0.1$, we have $[1 + 0.1(5.3 - 1)] = 1.43$ as the ratio of increase of variance over the first design. But this is considerably less than the ratio of 2.5 of the second design over the first. On the other hand, the cost of the third design is slightly greater than that of the second (75 blocks in contrast with 25) but perhaps in this case considerably less than that of the first.

Suppose that after examination of a few other designs (say, clusters of 2 and clusters of 8), the third design, of clusters of 5.3 dwellings, is adopted. The plan is then to select $1/10$ of the blocks and to select $1/3$ of the dwellings within the selected blocks. The probability of selection of any dwelling unit is $1/10 \times 1/3 = 1/30$ as before. Note that the inclusion in the sample of any dwelling depends on its being selected in two separate events. There are two stages of selection: first a sample of blocks is selected, then from the list of dwellings found within the sample blocks a sample of dwellings is subselected. This is an example of multistage sampling in general (see Section 21) (1, pp. 215-267, 3, pp. 135ff and 372-397).

At each stage of sampling there must exist a listing procedure for all the sampling units among which the selection will be made (Section 5). The listing of the blocks of the city is accomplished by the division of the map into blocks and the numbering of these blocks. After the sample blocks are selected, a listing must be made only of the dwellings in the sample blocks (15). The procedures of selection in the two stages are determined separately, and each of them may be a simple random selection or stratified, systematic, etc. In the selections of blocks and dwellings, as in many other undertakings, the use of systematic samples is quite usual. In our example the blocks would be selected with the interval 10 after a start with a random number from 1 to 10. In each sample block an interval of 3 would be applied to the listed dwellings after a random start from 1 to 3. Then we would obtain in each sample block a number of sample dwellings which is closely proportional to the total number of dwellings listed in the entire block in the proportion of 1 in 3. However, there would

be divergences of $1/3$ and $2/3$ of a dwelling from the exact fractional ratio

The numbering of the blocks may be used to introduce stratification. All that is necessary is to have relatively similar blocks (by geography or other variables) numbered consecutively. The interval applied to that list will select one block from each group of blocks equal in number to the interval—in this illustration, one from every ten.

The design has general applicability; similar considerations in other situations may lead to the same design of subsampling. For example, we might use an interval to select among all the work groups of a factory, subsampling with another interval for employees within the work groups.

20 *Subsampling with Probabilities Proportional to Size*

In Section 19, a sample design was described for a two stage sample of a city with an interval of 10 for selecting blocks and an interval of 3 for selecting dwellings within the selected sample blocks. The probability for selection of any dwelling in the city was $1/10 \times 1/3 = 1/30$. The procedure for selecting sample dwellings with the use of a constant interval within the blocks will yield in each sample block a number of sample dwellings which is closely proportional to the total number of dwellings existing in the block. But that number of dwellings will vary from block to block—sometimes a great deal. On the other hand for reasons of efficiency and convenience, we would like to keep the number of sample dwellings more nearly constant. However, were we to accomplish this simply by increasing the sampling interval within the larger blocks we would thereby decrease the probability of selection of the dwellings within these larger blocks and thereby violate our stated aim of equal probability (of 1 in 30) for all the dwelling units in the city.

We may remain true to that aim and also accomplish our purpose if we increase the probability of selection of any block in the same ratio that we decrease the probability of selection within that same block. Thus in the process of numbering blocks a block which appears to have about twice the usual number of dwellings is given two consecutive block unit numbers, thereby its chance of being selected is doubled from $1/10$ to $2/10$. When one of these blocks falls into the sample the interval of sampling within it is doubled—that is, every

sixth address is selected after a random start between 1 and 6. Hence, the probability of selection of a dwelling in that block is $2/10 \times 1/6 = 1/30$, as required (3, pp. 393-397).

The over-all probability of selection of any dwelling in any block is the product of the two probabilities: the probability of selection of the block, and the probability of selection of the sample dwelling within the sample block. During the process of numbering the blocks we assign block unit numbers as "measures of size" (Z_i) in accord with the number of apparent dwelling units in the block, and this measure Z_i is used in opposite and balancing ways in the two intervals pertaining to the two stages of selection. It is to the list of consecutive block unit numbers that the interval 10 is applied. The probabilities of selection are now represented for the two stages as

$$\frac{Z_i}{10} \times \frac{1}{3Z_i} = \frac{1}{30}$$

Here $1/30$ stands for the over-all probability of selection of any dwelling in any block in the city. The probability of selection within a specific (the i th) block is represented by $1/3Z_i$, that is, by an interval of $3Z_i$. The measure of size Z_i varies from block to block, for most blocks it is one, for some it is two, and for others three or more.

The probability of selection is determined by the application of the sampling intervals and is independent of any inaccuracies and deficiencies in the measure of size (Z_i). In so far as the measures of size are made proportional to the total number of dwellings in the block, the procedure yields equal numbers of subsampled dwellings from blocks which vary in total size. In practice this process cannot be complete, and it need not be. The subsample will vary somewhat, but it is enough if most of the variation can thus be eliminated (15).

Where does one obtain the information he needs on number of dwelling units in each of the blocks of the city so that he may assign measures of size? There are several sources of materials, one or more of which may be available for the city. Some of these are census block statistics, aerial photographs, the city engineer or the city planning office, the real estate board or the chamber of commerce, the local bank, the local newspaper, some local public utility, etc. Other sampling materials, as well as advice, may in some cases be obtained from the Census Bureau. If all these fail, one may resort to some cheap

rough estimates (perhaps just a glance from a moving car) to be made for all the blocks of the city or for a large sample of them

The use of block units (Z_i) is aimed at equalizing roughly the number of dwellings subsampled from blocks which contain unequal numbers of dwelling units. This procedure may be looked upon as an example of selection with "probabilities proportional to a measure of size." Let P_i denote the estimate of the number of dwelling units in the block. And let us say that we want to obtain an average of 5 of the (estimated) dwellings per block. Then we can take for our two stages of selection the probabilities of $P_i/150 \times 5/P_i = 1/30$. A procedure for selecting the blocks may be the following: (a) Assign measures of size P_i to each of the blocks. (b) Arrange the blocks in some desirable order, stratification may be obtained from the systematic selection because a block will be chosen from each successive sum of 150 of the P_i 's. (c) After a random start from 1 to 150, apply the interval of 150 to obtain the selection numbers. (d) Cumulate the measures of size P_i , taking them in the prearranged order. This cumulation can be done easily on an adding machine: whenever the addition of a block causes the sum to reach or pass one of the selection numbers, that block is selected into the sample. Each block thus has a $P_i/150$ chance of being chosen.

The procedure for selecting dwellings within the block involves the listing of all the dwellings in the block. Then the interval of $P_i/5$ is applied to the listing of the block, after a random start from 1 to $P_i/5$. Note that in order to avoid the inconveniences of an interval less than 1, in assigning them we should take no number smaller than 5 for P_i . Some special measures may be advised for the smaller blocks: they may be put in a special stratum, or they may be attached to the larger blocks, or they may simply be assigned a larger number arbitrarily. On the other hand, blocks so large that P_i is greater than 150 may be selected into the sample more than once. If a block is selected twice, then the interval for selecting dwellings within it should be halved to $P_i/10$.

The assigning of the varying measures of size will not eliminate all variation in the subsamples obtained from the various blocks. For one thing, the number of dwellings in the block will not be, in general, an exact multiple of the sampling interval. Therefore, the number of sample dwellings actually obtained may be different (by the fraction of one dwelling) from that expected on the average. Another source of

variation is more important it may be summarized as arising from the measures of size not being exactly proportional to the actual number of dwelling units listed in each of the blocks. This divergence may be due to a variety of reasons: inaccuracy or obsolescence of the source of the data, inaccuracies in listing, differences of definition, or differences in the units of measurement. All these possible differences will affect the actual size of the subsamples from the sample blocks. However, none of these change the probability of selection of any dwelling in any block because that is kept constant (at $1/30$) by the product of the two intervals.

This form of selection has general applicability. Let us suppose that we want to select a sample of employees of a factory made up of work groups which vary in size from 15 to 500 employees. Now we may want to satisfy two conditions: (a) To give every employee in the factory equal probability (say $1/30$) of being selected. Thus, the ordinary sample mean will be a proper estimate of the population mean for the factory. (b) To obtain a constant size sample (say 10 employees) from each selected section. Thus, the analysis dealing with sections as units will be facilitated and sometimes made more efficient. We have again an equation denoting the probabilities of selection in the two stages that looks like this:

$$\frac{P_i}{300} \times \frac{10}{P_i} = \frac{1}{30}.$$

P_i is the number of employees in the section or the estimated number in so far as exact numbers are not available. In the equation, the number $1/30$ satisfies requirement (a), the fraction $10/P_i$ satisfies requirement (b) either exactly or approximately, and $P_i/300$ for the selection of groups is necessary to make the equation hold.

The procedure will begin with the listing of all the sections. The order of listing may be utilized to yield stratification in connection with the systematic selection which will follow. Note that a section will be selected into the sample for each 300 employees. Cumulate the numbers of employees from section to section ending with the total number (12,000) in the population. Take a random number from 1 to 300 and apply the interval of 300. There will be $12,000/300 = 40$ selections made. But a section larger than 300 may fall into the sample more than once. If a section falls in twice, make a double selection from it.

The sample mean is again the simple $\sum x/n$. Incidentally, if in a sample of this kind 10 out of the 30 sections had a certain characteristic, it would be wrong to estimate from that datum that 1/3 of all sections in the factory were similarly characterized. Rather, one should estimate that 1/3 of all employees worked in sections which had that characteristic. The estimated variance of the sample mean is discussed in Section 18.

Selection with probabilities proportional to a measure of size is a widely used procedure. With its use the size of the subsamples can be stabilized in cases where the cluster size varies, in so far as useful (but not necessarily exact) measures of size are available. The probability of selection through the two stages is kept constant by the use of sampling rates, and it is not affected by the nature of the measures of size or by their inaccuracies.

This procedure is used in practice in two ways: first, intervals are used as shown in the two examples above; secondly, a single selection may be made from each stratum, as in Section 21.

21 *A Multistage Area Sample*

In the following description of the procedures for selecting the basic sampling units used for the national cross section samples of the Survey Research Center, some details have been omitted. It is presented as an example of the selection of a large scale multistage area sample. The set of procedures used in the several successive stages is basically the same. First, the population to be sampled is defined in terms of sampling units, and the sampling units are placed into strata and listed in each stratum. Then, the measures of size for all the units in the stratum are assigned and one unit is selected with probability proportional to its measure of size. Then, within each of the selected units there occurs a repetition of the processes, the defining, stratifying, and listing of units and the assignation of the measures of size and selection (7).

The design called for a sample of the 40 million private dwelling units* of the United States. We shall discuss a basic design adequate

*The sample of private dwellings excludes the residences of some people, chiefly hotels and institutions. They could be included by suitable procedures in a separate stratum. The 40 million figure came from the 1940 data used for drawing this sample.

for a sample of about 4000 dwellings, a sample in which every dwelling is to be given a $1/10,000$ chance of being selected. But it is understood that the same materials can be, and are, used over and over to yield various samples at different rates. Area sampling is used—the selection of area segments determines the selection of the dwellings associated with them.

The field work is done by part-time interviewers who cover, usually by automobile, an area within a reasonable radius of their homes. An interview "load" of about 50 sample dwellings per sample area (for one or two interviewers) was taken on the basis of considerations of the economy of clustering (Section 14). But 50 sample dwellings represent 500 000 dwellings in the population, enough to populate a state. So that the interviewer will not have to cover an area as large as a state, the sample area of each interviewer was to be restricted. The county was chosen as a basic unit providing an economic compromise between the aim of reducing the cost per sample dwelling and also spreading the sample over the diverse groupings in the population.

Most sample counties contain a city (usually the largest urban center in the county and near its center), several smaller incorporated cities and villages, some unincorporated congested areas mostly around the cities, and, surrounding all these, open country where farmers as well as many nonfarmers live. The central city is the most likely place to find a suitable interviewer. Usually half or more of the county's population (also half or more of its sample dwellings) are in or adjacent to the central city. The interviewer travels to the others (the outlying cities, villages, and open country segments) on roads which tend to radiate from her city. With an automobile she can go to one or more outlying sample points and return home the same day.

Thus, the county appears as a rather desirable sampling unit for the first stage of selection, as a "primary sampling unit." There are exceptions to the idealized picture drawn above. Some counties are too large or otherwise difficult to cover as an entity. These may be split into smaller units. In other cases, it is possible to build up from two or more counties, or from parts of them, primary sampling units which will have the desirable features mentioned above. From the point of view of increasing the spread of the sample over diverse areas, it is best to combine into one primary sampling unit contrasting kinds of areas. But to reduce the cost per sample dwelling, the amount of travel within the primary sampling unit is to be kept down.

and several other characteristics (7, 13) One particular stratum contains only three p s u 's, each with one of the large cities of the Southwest On the other hand, the county exhibited below is one of more than 100 p s u 's that compose this stratum, largely from the rural parts of the Plains States

The selection of the p s u was made from a list of all the p s u 's in the stratum, on which were cumulated the measures of size of those p s u 's (their 1940 populations) A random number from 1 to 1,831,704 was taken, since this was the total 1940 population of the p s u 's in the stratum The number chosen was one of the 19,277 numbers (between 253,613 and 272,889), any of which would have selected this particular p s u into the sample

The aim was to assign to every dwelling in the United States—hence also to every dwelling within each stratum—a probability of 1/10,000 of being selected From the stratum, a p s u was selected and its probability of selection was

$$\frac{19,277}{1,831,704} = \frac{1}{95\ 02}$$

What should be the rate of selection of dwellings within the county? It should be

$$\frac{1}{10,000} - \frac{1}{95\ 02} = \frac{1}{105\ 24}$$

because thereby the probability of selection of any dwelling in the stratum through the two processes of selection is made

$$\frac{1}{95\ 02} \times \frac{1}{105\ 24} = \frac{1}{10,000}$$

The requirements of probability could be met simply by listing all of the more than 5000 dwellings[†] in the p s u and applying the interval of 105 24 to it However, that procedure was judged to be too expensive Alternately, one could divide the entire area of the p s u, or county in this case, into small areas, segments in the open country and blocks in the towns, and apply the interval of 105 24 to those areas That

[†]We obtain these crude estimates of dwellings, as quick aids in planning the field work, using a factor of 3-1/2 or 4 persons per dwelling for the total of 19 277 people in the p s u

would be a feasible procedure in this largely rural county and suitable for some surveys. For our surveys it was judged that the town blocks with about 6 to 12 dwellings were larger than we desired. Therefore, it appeared that some subsampling was to be done within the p s u.

Maps, air photographs, and census data were procured to see what the county contained. It was decided that travel to a town was to be made for not less than 6 sample dwellings. This minimum for the sample yield in the towns (the secondary sampling units) is an economic consideration similar to the assigned yield of about 50 dwellings for the primary sampling units (Section 14). It means that a secondary stratum in the county should contain no less than roughly $6/50 = 12$ percent of the county. Thus the county was divided into the three strata given below with the 1940 populations for measures of size. Each of the strata is to be sampled with a rate of $1/105.24$. In

COMPOSITION OF A SAMPLE COUNTY

<i>Stratum number</i>	<i>Description of stratum</i>	<i>Measure of size (1940 pop.)</i>	<i>Cumulated measure of size</i>
I	Central city only urban place	3920	
II	All smaller incorporated places	872	872
		584	1456
		447	1903
		443	2346
		357	2703
		224	2927
		171	3098
		<hr/>	<hr/>
		3098	
III	Open country (the remainder)	12,259	
	Total for County	19,277	

dividing the county into strata and the strata into sampling units, the principles of listing must be observed: any area must appear in one and only one sampling unit (Sections 5 and 6). The division must be made clearly and consistently before selection takes place.

Stratum I consists of the one city. It was sampled in the two stages of blocks and dwellings, as described in Section 19. The product of the

probabilities of selection for the two stages was $1/17\ 54 \times 1/6 = 1/105\ 24$. Altogether, the probability of selection of a dwelling in that city depended on three selections

$$\frac{1}{95\ 02} \times \frac{1}{17\ 54} \times \frac{1}{6} = \frac{1}{10,000}$$

In Stratum II it was desired to restrict the expected yield of about 8 dwellings to one of the 7 villages which composed the stratum. The random draw from 1 to 3098 selected the fourth village with a probability of $443/3098 = 1/6\ 993$. In that town the rate of selection had to be made $1/105\ 24 - 1/6\ 993 = 1/15\ 05$. This was done again in the two stages for blocks and dwellings as $1/5\ 02 \times 1/3 = 1/15\ 05$. Thus a dwelling in that town was selected in four stages (county, town, block, dwelling), and the product of the probabilities is $1/95\ 02 \times 1/6\ 993 \times 1/5\ 02 \times 1/3 = 1/10,000$.

All of Stratum III, composed of the open country surrounding the other places, was divided into area segments.[†] There were 736 of these segments with an average of about 4 or 5 dwellings per segment. The interval of $1/105\ 24$ is applied directly to those segments, and every dwelling in the selected segments is included in the sample. Thus, about 7 segments, with about 30 dwellings in all, are expected from the open country. The selection of a dwelling in the open country is done in only two stages: $1/95\ 02 \times 1/105\ 24 = 1/10,000$.

The work performed in this county was similar in nature to work performed separately in each of 66 primary sampling units. The procedures were suited to the circumstances found in the primary sampling unit. It should be noted also that the expected numbers of sample dwellings are not "quotas." The selections are made with sampling rates, which are numbers controlled to yield sample dwellings with a fixed probability. In so far as the actual populations are different from the measures of size, there will be variations in the yield of the sample from different areas.

22 PROCEDURES OF ESTIMATION

We have examined in this chapter the role that some practical methods of selection can play in improving the sample design. The

[†]This division, as well as other useful materials, may be bought cheaply from the Census Bureau and is part of the *Master Sample*.

sample result with which we dealt (except in Sections 11 and 17) was the simple mean of the sample $\Sigma x/n$. But in all cases we understood that the sample mean is of interest only in so far as it serves as an estimate of the mean of the population which the sample represents.

Sample design was defined earlier as dealing jointly with procedures of selection and of estimation. The precision of the estimate based on a sample can often be improved—sometimes considerably—through the use of available auxiliary information. The topic is too complex for this chapter, but it is important, and the attention of the reader is called to it (1, pp 111-188, 22, pp 145-182).

For a simple illustration, let us say that the simple random sample of 400 out of 12,000 employees has been selected (as in Section 1). Suppose that we have the information on the proportions (N_h/N) of employees that belong in the several classes of some variable. Although we did not use that information for stratification during the selection process, after the data have been collected we want to use the information to improve the sample mean. The question may arise why was that information not used in the first place for stratification during the selection so as to obtain a proportionate stratified sample from each of those classes? Well, perhaps the sampler did not think of it at the time

		Total	Union	Nonunion
Number in class h	N_h	12 000	7200	4800
Proportion in class h	N_h/N	1 00	60	40
Numbers in the sample	n_h	400	260	140
Numbers of 'yes' answers	$p_h' n_h$	80	26	54
Proportion of 'yes' answers	p_h'	20	100	386

of selection. Or he had so many strata on the basis of other variables that he just could not use it for stratification in the selection process. Or, perhaps the data were not available beforehand.

Let us assume the last of these possible causes. Suppose that the union tells us that of all the 12,000 employees in the factory 7200 belong to the union. But union membership is not indicated on the payroll cards used for the selection of the sample. Suppose, also, that from the union, or from the respondents, we can find out who of the 400 in the sample belonged and who did not. It is necessary that the definition of belonging as ascertained for the sample cases be the same

as that used to classify the population into classes. The pertinent data on the weights for the two groups are given below, together with the results on the number of "yes" answers to a question from a simple random sample of 400 employees.

The proportion of union members in the population is $W_1 = N_1/N = 60/400$. With proportionate stratified selection, $60 \times 400 = 240$ of the sample of 400 would have been selected from these, actually 260 were. This is not a large deviation, but the chances of getting a larger deviation are only about 1 in 20.¹ Now the ordinary mean $p' = 80/400$ is, in effect, one where the strata are weighted in proportion to the representation in the sample. That is,

$$p' = \frac{260}{400} \times \frac{26}{260} + \frac{140}{400} \times \frac{54}{140} = 65 \times 100 + 35 \times 386 = 200$$

However, if we use the information we possess about the correct weights of the strata, we have

$$p_w' = 60 \times 100 + 40 \times 386 = 214$$

We see that the correction is not great in this case, particularly in light of the standard error of .02.

It is more productive to look at the matter in probability terms, as we did in the case of the proportionate sample in Section 9. The weighted mean p_w' in this case has a variance about 2 percent less than the variance of the unweighted mean p' . That is, by weighting the sample of 400 we make it as accurate as a sample of 408 would be without weighting. The gains in precision obtained by this kind of weighted estimate are just about the same as they would be by proportionate stratified selection procedure. This should not be surprising in

¹It must be noted that this last fact is in sharp contrast with the kind of adjustments by which conformity to population proportions is forced on data obtained in some haphazard fashion. The latter may have proportions grossly different from the population, hence the "corrections" may be very drastic. That means that the "corrected" mean may be very much closer to the population mean than the uncorrected mean. However, because the data do not arise from a probability sample, statistical theory will not bridge the gulf from sample mean to population mean. The conjecture must be made by expert judgment. It is even possible that the adjustment took the "corrected" mean farther from the desired population mean. Furthermore, no probability can be attached to the occurrence of this undesirable event.

view of the fact that the method of estimation in the two processes is the same, except for the differences in the number of actual cases obtained in the various strata. The gains are small, as in the proportionate sample of elements, and for the same reasons.

The above is but a simple illustration of a procedure for using auxiliary information in the estimation process in order to reduce the variance of the sample estimate. It is a poor illustration in that the gain in precision is very small. There are other estimation procedures applicable to other situations by which some large gains in precision may be made. The two most important are called ratio estimates and regression estimates. In some instances, the gains made through the use of better estimates are spectacular (1, pp 111-188, 3, p 87, 22, p 155).

BIBLIOGRAPHY

- 1 Cochran W G *Sampling techniques* New York Wiley, 1953
- 2 Deming W E Some criteria for judging the quality of surveys *J of Marketing* 1947 12, 145 157
- 3 ——— *Some theory of sampling* New York Wiley 1950
- 4 Dixon, W J and Massey, F J *Introduction to statistical analysis* New York McGraw Hill, 1951
- 5 Fisher, R A *Statistical methods for research workers*, 11th ed New York Hafner 1950
- 6 Goodman R Collapsed strata *Amer Statistician*, 1948 2, No 4, 22
- 7 ———, and Maccoby E E Sampling methods and sampling errors in surveys of the consumer finances *Int J Opin and Attitude Research*, 1947, 2, No 3 349 360
- 8 Hansen, M H, and Hurwitz W N The problem of non response of sample surveys *J Amer Stat Assoc*, 1946 41, 517 529
- 9 ——— Dependable samples for market surveys, *J of Marketing*, 1949, 14, 363 372
- 10 ——— Modern methods in the sampling of human populations *Amer J of Public Health*, 1951, 41, No 1, 662 668
- 11 ———, Marks E S, and Mauldin, W P Response errors in surveys *J Amer Stat Assoc*, 1951, 46, 147 190

- 12 ———, and Madow, W G *Sample survey methods and theory* New York Wiley, in press
- 13 Katona, G, Kish, L Lansing, J B, and Dent, J K Methods of the survey of consumer finances *Federal Reserve Bull*, 1950, 36, No 2 795-809
- 14 Kish, L A procedure for objective respondent selection within the household *J Amer Stat Assoc*, 1949 44, 380-387
- 15 ——— A two stage sample of a city *Amer Sociol Rev*, 1952, 17, 761,769
- 16 Lorie, J H, and Roberts H V *Basic methods of marketing research* New York McGraw Hill, 1951
- 17 Marks, E S Sampling in the revision of the Stanford Binet Scale *Psychol Bull*, 1947, 44, 413-434
- 18 McCarthy, P J Sampling elementary principles In Jahoda, M, Deutsch, M, and Cook, S W *Research methods in social relations Part II Selected techniques* New York Dryden, 1951, Chap 2
- 19 ———, and Stephan F F Area sampling *Amer Statistician*, 1951, 5, No 1, 20-21
- 20 McNemar, Q *Psychological statistics* New York Wiley, 1949
- 21 Sub Commission on Statistical Sampling of the Statistical Commission of the United Nations *The preparation of sampling survey reports* Statistical Papers Series C No 1 Lake Success 1949
- 22 Yates, F *Sampling methods for censuses and surveys* New York Hafner, 1949

PART III

Methods of Data Collection

In terms of a gross division, there are only three methods of obtaining data in social research: one can ask people questions, one can observe the behavior of persons, groups or organizations, and their products or outcomes, or one can utilize existing records or data already gathered for purposes other than one's own research. The last three chapters in this part describe techniques to be used in connection with each of these broad classes of methods. The first chapter is a theoretical analysis of the problems of measurement.

In social psychology, emphasis is placed upon methods of data collection in which the researcher himself introduces the specifications and the controls, and the chapters on interviewing and behavioral observation address themselves to this problem. The usefulness of existing records for many research purposes, however, is demonstrated in Chapter 7.

It is difficult to draw the line between what is and what is not methodology relevant to research in social psychology and related areas. Projective techniques, for example, ordinarily considered in the realm of "clinical psychology" have been used to good advantage on social psychological problems. Even techniques which belong in the realm of physiological psychology can be used in research on attitudes and group behavior. It was obviously neither desirable nor possible for this book to attempt to cover so wide a range of measurement techniques. The following chapters attempt to describe not all the methods which might conceivably be employed in social research but rather those methods of measurement which the person engaged in research in this area cannot avoid.

Problems of Objective Observation

*Helen Peak*¹

All scientists have as an ideal the objectification of their methods and techniques. That is to say, they aspire to observe, record, and interpret events in such a fashion that independent observers can verify their findings. The term "standardized measures" has been used rather narrowly by psychologists to apply to certain limited kinds of "tests" which are relatively specific as to procedures for administration, for scoring, and for interpretation of the scores on the basis of their reliability and validity, and with reference to established norms of performance. It is here maintained that whether one speaks of standardized tools in this restricted sense of the term or of objective measures with reference to all sorts of scientific observation, the problems are basically the same and revolve around the necessity of communicating concisely the conditions and procedures for making observations and interpreting them in the framework of a conceptual system.

This chapter will attempt to spell out certain fundamental

¹ I am grateful to Eugene Jacobson with whose collaboration this chapter was originally planned for a careful reading of this material and for his many useful suggestions. Many other colleagues have been helpful at various stages of the revision. I wish to mention particularly the editors of this volume and C. H. Coombs, D. W. Chapman, F. L. Kelly, and Earl Carlson.

questions that must be answered regarding instruments and procedures which purport to be objective, and will undertake to discuss some of the techniques that have been worked out to answer these questions. Space limitations have forced us to be selective in the choice of techniques which are evaluated and this selectivity has entailed the omission of many approaches which might well have been included. Furthermore, since many of the specific tools for the measurement and observation of social psychological variables are given detailed attention in other chapters of this volume, those used here for illustration in connection with answers to our basic questions have been chosen because they represent examples of important instruments not discussed elsewhere in the book. Although this criterion has resulted in an emphasis on individual rather than group processes, the questions discussed are relevant to all observational situations. The problems are general and basic whether we are dealing with group process and structure, with interaction, or with attitudes, motives, traits, and similar variables. Furthermore, they arise with respect to all sorts of procedures for observing these variables: interview techniques, participant observation, content analysis, questionnaires, rating scales, tests, and many others.

The problems to be discussed in this chapter will recur in many parts of a volume devoted to the analysis of scientific methodologies of social psychology. In so far as the other chapters are concerned with observational methods applicable to specific problem areas, however, they do not provide a general statement of the questions about observation and measurement. This is the task of the present discussion and of Chapter 11, in which Clyde H. Coombs has addressed himself to basic problems of measurement and observation.

THE PROBLEMS OF OBSERVATION

Whenever scientists set out to observe an event, some choice of instruments and procedures must be made. If instruments and techniques are ready at hand, questions will be asked about their characteristics and utility. If none are available which fit the pur-

pose, they must be constructed to specifications. In either case whether it be a matter of selecting an appropriate measurement device or of building one to order certain criteria are operating and certain questions are asked. To make clear what these questions are and to summarize and criticize some of the answers to them is our present task.

It will be convenient to organize the discussion around six basic problems

- 1 What *behavior* is to be *selected* and recorded in order to obtain the information required?
- 2 Under what *conditions* are observations to be made? How is the observational situation structured?
- 3 What evidence is there that some process with *functional unity* is being observed?
- 4 Has an attempt been made to summarize what is observed in quantitative terms? Can a *score* be assigned and what are the *metrical characteristics* of that score?
- 5 What is the nature and meaning of the process which has thus been observed or inferred? How is it to be labeled? What is its *validity*?
- 6 And finally, how stable are the observations? Can the same results be obtained under what appear to be the same conditions? Are the measures *reliable*?

The first two of these questions will be considered together in the following section, 'The Observational Situation and the Selection of Significant Behavior.' Questions 3 and 4 will be discussed together in the section on 'Evidence of Functional Unity,' because the demonstrations of functional unity are closely related to problems of scoring. Questions 5 and 6 will be treated in separate sections.

THE OBSERVATIONAL SITUATION AND THE SELECTION OF SIGNIFICANT BEHAVIOR

Something is to be observed and measured, and the investigator must start with hypotheses about the way in which this process or

event manifests itself in behavior, as well as the conditions under which such relevant behavior will appear. There may be occasions when interest lies wholly in seeing what people do (whether they cheat in a given situation, what answers they give to a question, what a group decision will be) with no desire to make inferences about traits or attitudes or processes. More often however, inferences will be made from behavior to variables such as tensions motives attitudes traits, perceptions or group norms which are only reflected in behavior. When this is so the investigator will have started with some theory (implicit or explicit) about the way in which these events affect behavior, and he will have chosen to observe those reactions regarded as appropriate for his purpose. In either case, it is the scientist's problem, if he is to be objective, to specify as clearly as possible what behavior or interaction patterns are to be noted.

Similarly, a statement of the standard observation situation must be made, including as complete a description as possible of the total structure of the situation for the reacting individual or group: what the observer does, the questions he asks or the stimulus items he presents, the instructions he gives, the comments he makes and the attitudes he reveals in the interaction situation. The inventory of conditions should also cover relevant facts about the reacting individual or social unit, particularly those psychological characteristics and processes which might affect the events under observation, and this will include that infinitely complex series of social variables which grow out of the interrelations of people with one another. It must be specified which of these are to be held constant and which allowed to vary in the standard observational situation.

In other words, if we assume that every behavior or reaction is a function of many antecedent conditions, the ideal of objective measurement requires the identification and observation of as many of the relevant antecedents of the event as are necessary to yield stable measures of it. This aspect of standardization has typically been neglected. There has been, for example, only tardy recognition of the fact that measures of intelligence have little meaning unless they are obtained under known conditions of education and motivation. As a rule, attention has been given to checking or controlling only limited aspects of the immediate situation such as the instruc-

tions to be given, time limits to be used, and the questions to be asked. Perhaps the failure to cope more adequately with other aspects of this problem stems from the implicit assumption that the processes being measured are independent of most variables and therefore relatively static and stable, an assumption which is clearly false for such processes as attitudes, needs, adjustment mechanisms, interaction, and other group phenomena. Here it is obviously just as necessary to have some notion about the existence of uncontrolled determinants as it is to know whether a patient has walked up a flight of stairs before taking a metabolism test.

Ideally, then, objective observation requires a clear indication of the behavior to be selected from the complex matrix of activity and definite specifications of the conditions for eliciting this behavior in such a way that inferences may be made about the variables involved. But it should be noted that there is no simple methodological prescription for meeting these requirements. They are the products of the theoretical sophistication, knowledge, originality, and hard work that are demanded in order to build theories about interrelationships between events and to test those theories empirically. To identify the important variables, to think of questions to ask, ways of structuring the situation, the appropriate behavior to observe—these steps depend to an important extent on the flash of insight and the hunch founded on knowledge and experience with the problems under consideration. The rules for the formulation of questions and statements, which are found in methodology texts, are simply convenient criteria against which to test ideas but have nothing to do with producing them. And it will be apparent that here, as at many other points, the theoretical model which the investigator brings to the task will play a crucial role, for it will be a major source of the ideas which occur to him and of the choices which he makes. If, for example, he sets out to devise a measure of hostility with a knowledge of the psychoanalytic theory of defense mechanisms, the questions asked and the behavior observed will be very different from that which would seem relevant if manifest expressions of hostility were regarded as the only appropriate data.

THE EVIDENCE OF FUNCTIONAL UNITY*The Meaning of Functional Unity*

To ask whether a test or series of observations has succeeded in isolating a characteristic or event with some sort of unity is to raise a question that observers of human activity and experience have been asking for a long time and in many circumstances. Over the years human nature and human behavior in its most inclusive sense have been sliced many ways but the resulting segments have often proved to be artifactual and without systematic significance. Wherever discriminations can be made on the basis of qualities of an event or experience or on the basis of its intensity and amount or in terms of differences in relationships or on any basis whatever a line may be drawn and a thing identified. So it is not surprising to find some 18 000 trait names in the well known list which Allport and Odbert assembled from English dictionaries (3). The ability to make discriminations is always involved in the isolation of parts or segments of reality but the reduction of this chaos of distinctions to scientifically useful concepts demands operational methods for discovering order and organization and for testing the meaning of the organizations discovered. We shall review and evaluate some of the procedures devised to determine the possible lines of fracture and foci of organization in the complicated matrix of behavioral events.

In the simplest case processes or events or objects may be regarded as having unity by virtue of sharing some common characteristic. Thus a group of round objects has unity with respect to roundness or you may classify responses in terms of a quantifiable characteristic such as speed or as leading to some consequence such as injury or support of another. If events share these characteristics they are in some sense homogeneous. The number of possible dimensions that may be described in terms of such essentially static similarities is legion. Such categories have their uses indeed they are indispensable.

But for the systematic purposes of science it is usually more fruitful to look for *functional unities* that indicate more than superficial similarities among events. To say that processes or behavior events are functionally organized has one of three meanings it may

mean that (1) they change concomitantly; (2) they are dynamically interdependent; or (3) one process changes dependently with another (cause and effect). Most of the methods for discovering the presence of functional unity deal simply with concomitant change; *i.e.*, they derive in large part from correlational techniques and consequently are mute with respect to the existence of dynamic interdependence or causal relations. These distinctions between kinds of functional unity will be discussed further in the section on "Dynamic Organization."

The first task is, then, to describe and evaluate certain of the correlational procedures for demonstrating what is here called functional unity. This is not to inquire about *what* is being measured but rather about the existence of some process or some aspect of an event with sufficient integrity that it may be identified as organized.

The Problem of Scoring

As we examine the concept of functional unity, it will become clear that it is intimately related to the question of scoring. Some method of assembling the evidence from multiple items of behavior into a composite is needed in order to reduce the raw data to manageable form, and scoring is one form of summarizing. It is possible, of course, in the simplest case, that the data will be a correct or incorrect answer to a single task, an agreement or disagreement with a single statement about an isolated issue, or simply the fact that a child either got into a fight or did not under some condition. Typically, however, many behavior items will have been noted, and these must somehow be drawn together. This may be done intuitively and the result expressed verbally without benefit of numbers, but it is clear that where procedures for combining data into scores are definitely prescribed, errors of subjective interpretation and bias are eliminated in some degree. Furthermore, there are many obvious advantages in numerical indices, always provided they do not distort the data. These problems will be discussed below.

In any scoring procedure, several steps are involved, and the first of these requires that the items of behavior which are to be combined into a score must be classifiable into the same category on some grounds. This may involve simply an *a priori* decision that all the items in a test do, in fact, fall in the same category—

that for example, they all deal with Russia, or with extroversion, or with achievement. Some statistical or experimental method of determining functional unity may have been employed to give empirical evidence of a common factor or dimension or characteristic. In any case, if there are many separate events which are to be reduced to more simple form, an inescapable step will be *categorization*. This simple mapping of objects into symbols" results in what Coombs and others have called the nominal scale, though it constitutes a step in building any sort of scale.

At this point there will usually be some resort to numbers in order to express the observational results as succinctly as possible. This typically involves the process of *counting* the behavior items which fall into a specific category: the number of items passed or agreed with, the number of instances of sympathy, or fighting, or defensiveness, the number of judgments that *A* is better than *B*. There are other possible quantitative measures, such as speed, duration, extent and intensity of categorized events, but these have not been used as widely in social psychology and, although important, do not present the same order of problem.

But there are questions about the rationale of counting. It must be specified for example how each behavior item is to be *weighted* in the counting process which yields the total score. Do we assign one point to each statement agreed with or shall we weight some items more heavily than others? Should different weights be given to statements assented to with different intensities? The problems of weighting will be considered in connection with the procedures for determining functional unity (see pp. 252, 253, 256, 262, 273, 275). Suffice it to say at this point, that the major problem lies in finding a theoretically meaningful rationale for determining that items shall be given the same or different weights in the composite score, and this amounts to asking whether a way can be found to discover how the processes reflected in the item answers combine to produce their effects.

Then there are the many questions about the *metrical significance* and uses of scores. When we have calculated a score, what meaning does it have as a number? Do equal steps between scores represent equal distances on a psychological dimension? Is there an absolute zero? What sort of mathematical operations can we

use with these scores? How do we best express the relative scores of persons when comparing them with others?

The answers to these questions open up the important problem of the relationship between the model implied in the use of mathematical operations such as addition, multiplication, correlation, etc., and the data to which the operations are applied. Coombs, in Chapter 11, has described the characteristics of various types of scales, and the reader is referred to that discussion in this connection. The important point is this: If we impose on a set of data assumptions, such as the common one that equal steps between numbers assigned to reactions represent equal intervals or increments of some psychological process, the results that emerge from the mathematical operations may be determined more by this assumption than by the nature of the reality that is being measured. If we assume unjustifiably that our scale positions are equal psychologically and proceed to manipulate the numbers in accordance with these assumptions of continuity and equality, the resulting means and correlations will not reflect an accurate picture of the processes inferred to exist.

Specific Procedures for Determining Functional Unity

Certain operational procedures will be discussed in this section from the point of view of the evidence of functional unity which they provide. These procedures differ in a number of respects, including the level of behavior complexity to which they are adapted. Those to be discussed first look for unity and organization among relatively simple behavior items, such as responses to single questions or statements. Here we consider the analysis of internal consistency among items by item analysis and by the demonstration of some rational order among behavior items by the scaling techniques. At the next level the units or segments of behavior are total test scores; correlations between tests are studied for evidences of organization. At a more complex level factor analysis seeks simplifying relationships among intertest correlations. Finally, certain special problems of establishing dynamic unities will be discussed, these are typically complex organizations.

For each of the methods of analyzing functional unity, we shall

raise the following questions (1) What, briefly, is the method? How are scores arrived at and how is functional unity determined? (2) What assumptions and limitations are implicit in the use of the procedure? (3) In what sense does the method discover functional unity?

ANALYSIS OF INTERNAL CONSISTENCY BY ITEM ANALYSIS Although item analysis² has been used most often as a basis for selecting test items in the interest of improved prediction to a criterion, it may also be thought of as a device for establishing the existence of functional unity within a test. In the simplest case, the procedure involves finding correlations between test items, using such statistical tools as tetrachoric r and the phi coefficient and selecting for the final form of the test those items with highest average correlations.³ Closely related is the procedure of finding the relationship between each test item and total test score. This may involve correlating item response and test score and determining the regression of total score on the item (*i.e.*, the prediction of total score from item response) or the regression of item on the total score (prediction of item response from total score) (1)

Any one of the correlation methods in which at least one variable can be expressed in dichotomies may be employed for determining item test correlations (*viz.*, tetrachoric r , phi coefficient, biserial r , point biserial), provided the assumptions involved in the use of these procedures are met (24, Chap. 13). But before a total test score can be calculated from item responses, a decision must be made about how items are to be weighted and combined to produce the score. Often weights are assigned arbitrarily and without any rationale. Thus any item agreed with or passed may be given the same weight in the total score as any other item agreed with. This has typically been the procedure for attitude tests. The Likert method (31) is a special instance of this procedure. In responding to the statements which make up a test, individuals are asked to indicate the degree of agreement or disagreement with each item. If five steps between extreme agreement and disagreement

² The reader is referred to standard texts for a detailed discussion of the techniques of item analysis (25, Chap. 21)

³ See Loevinger (32) for a discussion of selection of test items for purposes of prediction to a criterion which involves choosing items with low intercorrelations and high correlations with a criterion (sampling principle) as compared with selection looking to valid self consistency of items (equivalence principle)

are used, weights ranging from one through five are assigned, so that extreme agreement with one item gets the same weight as extreme agreement with another item. The score is the sum of the weights.

Many other methods of assigning weights have been worked out (25, Chap. 18) (28, Chap. 7, Supp. Study B-4 and D). For example, weights may be determined on the basis of the correlation of each item with some criterion to which the test is designed to predict. This often means that the weights must be changed when predicting to different criteria. Or weights may be selected so that the dispersion of the individual scores is a maximum. But the difficulty with these and many other methods is that they do not commonly attempt to offer any theoretical reason concerning the manner in which the processes reflected in the items actually work together to produce the predicted consequences.

These correlation techniques have serious inadequacies as tests of functional unity among the items of a test, whether they are applied to the analysis of relationships between items or between items and total score. In the first place, some of the methods run into problems regarding statistical assumptions. Indices, like tetrachoric and biserial r , assume linearity of the relation between variables, normal distribution, and continuous variables. It is impossible in some cases to determine whether the data conform to the statistical assumptions—for example, the assumption of continuity in dichotomously answered item variables. In other cases the requirements may be clearly inappropriate to the data, as when tetrachoric r is calculated from tables having a single cell with a zero entry.⁴

In the second place, the interpretation of these indices as indicators of the degree of functional unity among the items of a given instrument is often ambiguous. There is, for example, no satisfactory way to disentangle the combined effects on correlation of the presence of items at various levels of difficulty and the degree of heterogeneity of processes involved in responding to the items of the test. Some indices, such as ϕ and point biserial, have artificial limits set on their size by the difficulty of the items correlated. The maximum value of ϕ , for example, is unity only when the same proportion of people pass both items involved in the correlation. But it is

⁴ See Loewinger (32 pp. 17ff.) for a discussion of these indices of item validity.

necessary that a discriminating test include items over a range of difficulty. In this case the index gives little information about the degree of functional unity between items or between an item and a total score (24, p. 342).

The size of the correlation coefficient between an item and total score will be affected by still another factor, the length of the test. Obviously the relation of a single item and the total score will be closer for the short test than for the long test and will tend to decrease as test length increases. Gulliksen suggests that this tendency will be reversed when the point is reached at which the added test length increases the reliability of the total score so much that an increase in the item test correlation will result. The relationship over such a range has not been investigated (25, p. 393).

In interpreting the meaning of these correlations, we must also remember that no one of them can be regarded as the equivalent of a Pearson r , unless the assumptions underlying that coefficient have been met. This being true, the apparent advantages of phi and point biserial vanish. Although they can be used with variables that are not continuous and not linearly related, the meaning of an index of a given size in these circumstances is highly ambiguous.

As already indicated, the item-total score relationship may also be expressed as the regression of total score on an item or of an item on the total score. An illustration of this procedure is found in what is sometimes called the index of discriminatory power (35). The total scores of those who answer a specific item in different ways are compared to see whether those who respond to the item favorably have a total score falling at the favorable end of the distribution and vice versa. The procedure may, of course, be reversed by observing the average item scores of those at the extremes of the distribution of total scores. In either case, when the prediction succeeds, some relationship has been shown to exist between the item and the score, in so far as this can be demonstrated by plotting two extreme points on any distribution. One limitation of this procedure for selecting related items is that only part of the data is used, making it possible that an apparent relation between item and total score will disappear when the middle parts of the distribution of scores are included in the analysis.

It must be concluded that these methods usually yield unsatis-

factory and ambiguous evidence regarding the existence of functional unity among the items of a test. Each item may be correlated to some extent with every other item and with the total score but this still does not mean that all the items involve the same process, even to some limited extent. The existence of correlation between pairs of items is a necessary but not a sufficient condition of the existence of unity between all the items that make up a test. Finally, nothing is revealed about the nature and complexity of the processes measured by the instrument made up of items thus selected. There is only evidence that some of the items are related, but because of statistical ambiguities the extent of this relationship is uncertain. At best, the methods are appropriate to the initial stages of analysis of this problem of functional unity or to the selection of items of a test to be used chiefly for prediction to a specific external criterion.

ANALYSIS OF INTERNAL CONSISTENCY BY RATIONAL ORDERING OF ITEMS. Here, as in the preceding section, we shall be concerned with item analysis. But the scaling methods go beyond correlation of items and make the attempt to order the items with respect to one another—*i. e.*, to assign scale positions to them. Certain of the scaling methods proposed by Thurstone, Guttman, and Loewinger will be examined in order to determine the contributions of these techniques to the solution of the problem of functional unity.⁵

In the construction of Thurstone's equal appearing interval scales the selection of items and the process of scaling proceed simultaneously. Many statements about a given subject are prepared, and judges are instructed to sort the statements into eleven piles considered equal distances apart and ranging from most favorable statements at one end to least favorable at the other end with a neutral position in the middle. On the basis of these sorts, it is possible to determine the mean or median position assigned to each statement by the various judges and the interquartile range (*Q*) of the assigned positions. Statements with low *Q* values and with median scale positions approximately equal distances apart are

⁵ These have been chosen as types of scaling procedure that raise most of the important points to be considered. The most important procedure which has been omitted is Coombs' ordered and partially ordered metric based on multiple comparisons which is discussed in Chapter 11 (see also 11). Edwards and Kilpatrick have suggested an ingenious combination of the Thurstone Likert and Guttman procedures (16).

selected for the scale. Persons who take the test check those items with which they agree and the test score is the scale value of the median item checked.

Thurstone considered it appropriate to use means and standard deviations with scores obtained on these tests because he regarded them as interval scales (41, 43). This implies, of course, that numerically equal changes in scale position at any point in the scale may be regarded as indicating equal amounts of change in attitude. It must be remembered, however, that the status of this assumption remains doubtful for a number of reasons. Even though judges consider the eleven piles as being equal distances apart, this does not mean that the processes inferred from the statements actually change linearly with these perceived equal distances. This places on the unsophisticated judge the responsibility for interpreting not only the relative amount of unfavorableness implied by agreement with certain statements but also the increments of unfavorableness. In other words, he must do without any knowledge of the problems involved what the experts have found impossible.

It is essential, at the very least, that there should be a convincing theoretical justification for making any assumption about the relation of an inferred process to the phenotypic response, and such theory is wholly lacking. Moreover, there is inevitably distortion in the scores obtained by the persons who fall at the extreme ends of the scale because of the restriction of range of items at these points. A score of eleven (maximum) could be obtained only by agreeing with the one most extreme item. If any other items are checked, the median will be pulled in toward the middle of the scale. It is also likely that there is distortion in these extreme intervals by the judges themselves due to what has been known in psychophysics as the end effect. This is the tendency for judgments to pile up in the end categories (23) and distributions of statements by the Thurstone sorting method suggest that this actually happens. Assumptions about equal intervals should therefore be regarded with skepticism.

The distinguishing characteristic of this scale is then that judges and not respondents determine the meaning and the ordering of items and consequently the value assigned to the inferred variable. This involves the assumption that there is some common dimension to the statements judged and that this dimension can be abstracted. The dimension may be favorableness toward an object

as in the attitude tests, or excellence of handwriting, or amount of threat implied in certain actions, or cohesiveness of a group, or anything else. The evidence that the task is a feasible one and that some such dimension does exist rests largely on the fact that substantial agreement among judges can be reached regarding the relative scale position of items. Moreover, a number of studies have seemed to show that statements scaled by the Thurstone method maintain their scale positions even when persons with widely differing attitudes serve as judges (19, 27, 36). But as Carter (7) has indicated, this question cannot be answered once and for all, and it is necessary to take the problem of sampling into account when choosing judges to be used in the standardization process.

Furthermore, scale positions have been shown to be affected by other conditions. Farnsworth, for example, found marked shifts in position of items on the Peterson Attitude Toward War Scale from 1930 to 1940 (17). More recently, Hovland and Sherif (30, 37) have pointed out the inconsistency between the two sets of findings (1) that assigned scale positions and judges' attitudes are independent, and (2) that perception and judgment are affected by motivational and attitude factors. They have reopened the entire question by an investigation of the influence of judges' attitudes on the positions and distributions of items sorted by the Thurstone method. The judges used were Negroes, pro Negro whites and anti Negro white groups. There is evidence of marked piling up of statements in unfavorable categories by Negroes and pro Negro whites and in the pro Negro categories by anti Negro whites. It appears, however, that even in this experiment there is no grave distortion of relative item positions. The implied equality of units is meaningless and perceived distances between items are greatly altered by the judges' own attitudes, but roughly the same items still appear to be more favorable or less favorable than others to all the groups. The ρ between item rank assigned by Negroes and anti Negro white subjects is 0.937.

The question inevitably arises as to whether individuals work within the same frame of reference when they take the test and whether respondent reactions give evidence of the same ordering of items as do the reactions of judges. Thurstone himself was interested in this problem and in the late 'twenties devised what he called the index of similarity or relevance, based on logic which is

essentially the same as that adopted more recently by Guttman, Loevinger, and others. He sought evidence of how respondent answers were patterned and whether persons agreeing with a statement in a given scale position accepted only statements in adjacent positions or whether they checked items scattered over a wide range of scale values with gaps between. If the items are perfectly ordered by the respondent, there should be no reversals in judgment (no items disagreed with) between the first and the last items agreed with. This is the basic logic of scaling when there is a single response to each item by each person. The index of similarity was designed to give some notion of how each item fell into such a pattern (41, 43).⁶

After statements of known scale position and low Q values have been selected for a scale by Thurstone's method, the test is administered to a group of subjects who indicate the items with which they agree and disagree. From these data the index of item similarity is calculated. The index for any item A with respect to another item B is based on the total number who endorse item B (n_b) and the total number who endorse both A and B (n_{ab}). The index is the ratio of n_{ab}/n_b . If everyone who endorses B also endorses A , then the index is unity. If no one endorses both A and B , the index is zero. After these indices have been determined for all items in relation to A (the item being tested), they are plotted against the scale values of the various items. The similarity of the item A with respect to other items is estimated from the appearance of the whole diagram. If the indices of statements nearest in scale value to the test item A are high and if they drop off in both directions as we move away from this item A , then A is regarded as satisfactory because it has been shown that people who endorse this statement are not likely to agree with statements which are at distant points on the scale. Although the calculation of the index for each item in relation to every other item is clear cut and unambiguous, the decision whether a given item is suffi-

⁶ When a test is of the cumulative type, it scales perfectly if answering correctly or agreeing with a more difficult item means that all easier or more acceptable items have been answered correctly or agreed with. In the differential type of test, a person in the middle of the scale will disagree with all the less favorable items, agree with a series of items of moderate degrees of favorableness, and disagree with all extremely favorable items. There will be no gaps. The terms cumulative and differential are Loevinger's (33). Thurstone uses the terms increasing probability type test and maximum probability type test (41). Coombs somewhat more general terms monotonic and non monotonic refer to this same distinction (see Chap. 11).

ciently relevant to belong in the scale must be a matter of judgment. Moreover, there must be as many graphs as there are items, and no method is provided for representing the homogeneity of the test as a whole.

It appears that Thurstone has not made extensive use of this technique in the construction of his scales. Edwards reports a communication from Thurstone in reply to an inquiry on this point which states that he has come to feel that items should be selected on the basis of factor analysis rather than by the criterion of similarity (15).

In any case, it is clear from a number of studies that the Thurstone scales, when tested by the index of similarity, often contain statements which do not maintain the scale position assigned by judges. Edwards (15) has shown, for example, that the neutral items in a number of Thurstone scales are likely to be accepted by persons who agree with statements widely scattered on the scale. Dudycha produces similar evidence in a study which reports the analysis of the scatter of endorsed statements on the Peterson Attitude Toward War Scale (14). He found that individual students with highest and lowest scores often showed endorsements of statements ranging from strongly favorable to strongly opposed, and he questions the value of medians derived from such widely scattered scale values.

The fact is, then, that the functional unit abstracted out by the Thurstone method of equal appearing intervals is not usually the same as that which is found by scaling techniques based on respondent reactions, and the question will be asked whether the organization found by one or the other of these methods is the more truly fundamental. The answer must be given in terms of the efficiency of the constructs which the measures yield for specific purposes. If, for some purposes, constructs based on scales which are ordered in terms of respondent answers produce more verifiable predictions to behavior or to other constructs, such scales can be regarded as superior for these purposes. The possibility exists that judged scale position of statements may prove to be useful for quite different purposes, many of which have not been adequately explored. Such consistently ordered series of items might, for example, be regarded as measures of group norms regarding the implications of certain statements rather than as a basis for accurately inferring an attitude.

That is to say, consensus about the meaning of behavior may thus be determined. Such information has potential usefulness in the interaction situation where, rightly or wrongly, we make such inferences and interpretations. Similarly, such scaling devices are appropriate for determining the generally anticipated social sanctions following certain actions, information that is potentially of great importance to social psychology.

A final point must be raised. Heterogeneous measures such as Thurstone's or Likert's (31), which do not scale in the Guttman sense of the word, will result in a situation in which equal scores are obtained by persons who accept widely different items. In this case the pattern of item responses cannot be predicted from knowledge of the total score. This is usually interpreted as meaning that the scores do not represent a measure of the same process. This may be true but, on the other hand, it must be remembered that the same behavior may be determined by different patterns of antecedent conditions. In this case there may emerge from reactions to an apparently heterogeneous collection of items of a scale an index of some process which reflects a tendency to positive or negative action and two persons agreeing with different statements may in fact be comparable in the strength and direction of this tendency to act in a given direction. We do not know as yet that independent measures of the same unidimensional processes in different people furnish the best means to the prediction of behavior.

The method of analysis by Guttman's unidimensional scales has taken as its central problem the discovery of unidimensionality in a sample of test items drawn from a universe whose content is defined and described arbitrarily by the investigator or by a group of judges. The universe may be marital adjustment, opinion about British fighting ability, the esthetic qualities of paintings or any other area of interest and it is assumed that the ordering of persons based on a sample of items will be essentially the same as that based on the whole universe of items (40, p. 81). The evidence for the presence of a scale is found in respondent reactions to a sample of items rather than in the opinions of judges about the scale position of items and the essential meaning of unidimensionality is the operational fact that respondent answers to test items order themselves in such a way that all the item responses can be

reproduced from knowledge of the total test score. In the cumulative type of test, this will be the case when it can be shown that any person who agrees with a given item will agree with all less extreme items and will disagree with all more extreme items. The same score cannot be made by persons who accept different statements. For the differential type of scale, respondents will accept a group of items which are adjacent in scale position, disagreeing with those statements on either side of this position.⁷

The details of Guttman's method are described in Volume IV of the American Soldier Series (40). Stated briefly, the procedure involves mapping respondent answers to each item onto a scalogram in which items are placed horizontally across the top of the table, ranging from those most favorable to those least favorable. Respondents are ranked on the vertical margin of the table, with individuals giving the largest number of favorable responses at the top. After these preliminary steps, items and respondents are shifted in position until the maximum order possible with the given set of responses is discovered. The ideal sought is a pattern in which any person *A* with a higher total score than any other person *B* will have agreed with items as high as or higher than *B*. If there is no scale, items agreed with will scatter all over the scalogram without relation to the rank order of the total scores and no rearrangement of items or persons can reveal this kind of order.

The extent to which a given series of items departs from the ideal of unidimensionality is expressed by the coefficient of reproducibility. This index is simply the proportion of item responses that can be correctly predicted from knowledge of the scale scores of persons taking the test. When the scale consists of 5 items which have been given to 100 people, the total number of responses will be 500. If 50 errors are made, the coefficient of reproducibility will be 0.90, which is set as the minimum value permissible for a "scale" (40, p. 77).

Although this is the principal criterion for the existence of unidimensionality, Guttman suggests that at least four other factors should be taken into account in making a judgment on this matter. Two of these should be mentioned briefly at this time. The items of the test should cover as wide a range of marginal distributions as possible, and items showing a 50-50 distribution (*i.e.*, 50 percent

⁷ See footnote 6, p. 258, on cumulative- and differential-type tests

of respondents agree and 50 percent disagree) should be included. This is recommended in order to avoid the spuriously high coefficients which result from the fact that the reproducibility of any one item can never be less than the percentage of respondents falling into the modal category (*i.e.*, the category in which the most people are found). If all test items show a high proportion agreeing or disagreeing, the coefficient will inevitably be high regardless of the internal organization of the items. This amounts to saying that the size of the coefficient is dependent to some extent on the *difficulty* of the items of the test.

Guttman also says that attention should be paid to the pattern of the errors (*i.e.*, to those cases where item responses are incorrectly predicted from the total scores) in order to distinguish random errors, which produce quasi scales, from those errors which indicate that there is no order among the items (nonscale types). This judgment becomes the basis of selecting items which scale and of rejecting those judged to belong on other dimensions.

One of the advantages of this method is that problems of assigning weights to each item of a scalable test are minimal. This follows from the simple fact that the relative scores obtained by individuals would be the same regardless of the absolute size of the weights used, provided the weights correspond to the rank order of the items on the scale. When rank orders are used as weights, the score may be expressed equally well in a number of ways as the median rank of the items accepted, as the weight of the most favorable or least favorable item (in the case of the differential test), or as a sum of the weights. The relative scores would be the same in all these cases.

The Guttman scale is an ordinal scale, and no claims are made for it as an equal interval scale. Guttman does propose to determine a "zero point" which separates the unfavorable end of the scale from the favorable end by means of what he calls the intensity function. After the respondent has answered a given item on the test, he is asked "How strongly do you feel about this opinion: not at all strongly, not very strongly, fairly strongly, very strongly?" Scale scores are then plotted against the intensity scores thus obtained, and in a good many cases a U shaped curve results. That is, those who feel strongly tend to be those who fall at the extremes of the scale. The scale position corresponding to the lowest point

reached on the intensity dimension is regarded as the zero point which divides the favorable from the unfavorable side of the scale.

Two comments should be made about this function. First, it is not altogether clear that it can establish a division between favorable and unfavorable attitude responses in all cases, for not all data show this U-shaped function and those which do typically have a great deal of scatter about each of the points—*i.e.*, many people falling at the middle or neutral point of the scale will feel strongly. Second, even when there is a U-shaped curve, the zero point cannot be regarded as an absolute zero where there is a complete lack of attitude or of affect toward the object of the scale. The determination of such a zero point, like the discovery of equal intervals, must involve some theory about the way in which the variable being measured is reflected in the observed behavior. There is no such adequate theoretical rationale as yet.

It is assumed by Guttman that "the invariant cutting point between favorable and unfavorable responses" can be determined by this method and that this point will be independent of the bias resulting from the specific questions asked. In other words, it is asserted that the percentage of persons favorable toward the issue or object will remain the same from one set of questions to another when the point of shift from favorable to unfavorable is determined by selecting the score corresponding to the lowest point on the intensity scale. Guttman provides some evidence that this is true, but it remains to be seen whether this location of zero has any validity in the sense of being related to a point of shift from favorable to unfavorable action.

The Guttman method of scaling can be evaluated from a number of points of view. First, questions have been raised regarding the ambiguity of certain steps in the procedure. We have indicated that Guttman recommends the application of several criteria in addition to the coefficient of reproducibility in judging the scalability of items. It has been pointed out that the application of these criteria is essentially subjective and that there are no clearly prescribed rules for arriving at a final judgment. For example, it must be difficult in practice to determine merely by inspection whether errors are random or systematic, yet this judgment is one step in deciding whether the data reveal the presence of a scale, a quasi scale, or a nonscale type of test. Likewise, the problem of the

relation of the size of the coefficient of reproducibility to the difficulty of the items (as shown by marginal values) is handled by the highly ambiguous recommendation that a wide range of item difficulties should be employed and that a number of items in the middle range should be included. It is not specified precisely how the coefficient may be interpreted with different distributions of item difficulty.

Another source of ambiguity is found in the procedure involved in determining the presence of order among items on the scalogram. There is no clearly specified method to follow in shifting columns and rows until the order is attained which is known to yield the maximum coefficient of reproducibility. Since it is impossible, in practice, to try every combination of rows and columns, a decision that the best possible arrangement has been found must be in some measure intuitive.

The fact that the size of the coefficient of reproducibility will be a function of the specific items selected at the outset poses another important problem that has not been solved. The size of the coefficient will be changed not only by the presence or absence of unidimensionality in the area but also by the fineness or grossness of the steps between items. Other things equal, when these steps are large there will be fewer errors of prediction of item score from total score and consequently a higher coefficient of reproducibility than when the steps are fine. As already indicated, Edwards and Kilpatrick have suggested the use of the Thurstone technique as a basis for the initial choice of statements which will have objectively determined distances between them (16). The Hovland-Sherif study reported above throws doubt on the usefulness of this procedure for such a purpose since the distribution of scale items is shown to be changed radically as a function of judges' attitudes toward the issue (30, 37). If, however, the judging group is chosen as representative of the sample on which the scale is to be tested and used, the Thurstone method might at least provide an objectively defined distribution of items which could be described as a condition of the obtained coefficient of reproducibility. It is clear that the problem is not unique to the Guttman method but it does point up another necessary reservation in interpreting the allegedly objective index, the coefficient of reproducibility.

In interpreting the meaning of unidimensionality, Guttman

indicates that scale analysis affords a rigorous test of the existence of a single meaning for an attitude area in the eyes of the respondents. This is an acceptable interpretation but Guttman goes on to make the further assumption that when a scale is shown to exist among items, the universe of which the items are a sample will also scale, as will any other set of items from that universe. On logical grounds this statement must also be accepted, but this is not the same as saying that we can in practice identify that universe from which scalable items come. Guttman says that a universe must be defined in this fashion:

An attribute or item belongs to the universe by virtue of its content. The investigator indicates the content of interest by the title he chooses for the universe and all attributes with that content belong in the universe. The evaluation of the content thus far remains a matter that may be decided by consensus of judges or by some other means (40 p. 84).

He points out, further, that samples from many universes so defined do not scale or are made up by several subscales. Items which scale under one set of conditions and with one population do not necessarily prove scalable in other circumstances. It appears, therefore, that the most that can be learned from the demonstration of unidimensionality is that a given set of items scales under the specified conditions and with the specified sample. Having determined this, we have no way of inferring directly what the universe is of which these items are a sample. Only empirical test of some theory about the reasons and conditions for the appearance of order among the items will make it possible to identify other situations (items, frames of reference, populations) to which one may generalize the findings.

The question may be raised whether the method is adequate for discovering all kinds of fundamental psychological variables. It may be appropriate mainly to the task of finding unidimensionality where answers to certain questions imply definite answers to other questions which are obviously related. For example, if a person is asked whether he wants to stay in the Army for a career after the war and answers, "Yes," it is reasonable to expect that he would also give an affirmative answer to the question "Are there any conditions you can think of under which you might consider staying in the Army after the war?" In such questions the dimension

reflected may be simply the respondents need to be logical. Although not all the unidimensional scales reported by Guttman are made up of such manifestly related questions, this does tend to be the case. How much beyond the discovery of obvious logical patterns this method of scaling can go in practice is yet to be determined.

Another serious limitation of the method must be mentioned. Coombs has pointed out that if a test is made up of items which scale in Guttman fashion, the opportunity to discover deviant responses will have been eliminated (10). This is another way of saying that perfect scales in the Guttman sense will be found only for psychological processes which are common to a population. Areas yielding unique patterns tend to be eliminated in the scaling procedure.

The logic behind Loevinger's technique of homogeneous tests is essentially that which Guttman employs. It assumes that on a perfectly homogeneous cumulative type of test a person who passes the more difficult item or agrees with the less popular item (accepted by fewer people) should also pass the easier task or agree with the more popular statement. The method makes no assumptions about magnitudes and is appropriate for qualitative data. Its use is restricted to the cumulative type of test. The essential difference in the two techniques is found, then, not in the logic of scaling but in the indices proposed for indicating the amount of homogeneity⁸ and in the interpretations of the meaning of homogeneity and unidimensionality.

Three indices are proposed for describing homogeneity: homogeneity of two items (H_{ij}), test homogeneity (H_t) and item test homogeneity (H_{it}) (33). Where p_i is the probability of passing the i th item and $p_{i/j}$ is the probability of passing the i th item among those passing the j th item (j is a more difficult item than i), then

$$H_{ij} = \frac{p_{i/j} - p_i}{1 - p_i}$$

If there is perfect homogeneity and all those who pass the more difficult item pass the easier one, then H_{ij} is equal to 1. When there is complete heterogeneity, it equals 0. Homogeneity over the whole

⁸ Loevinger prefers the term homogeneity to 'scaling' and uses it in much the same sense as Guttman uses unidimensionality (33).

test (H_i) is a weighted average of the H_{ij} 's for each pair of items adjusted so that the coefficient will equal 0 on the perfectly heterogeneous test and equal 1 for the perfectly homogeneous test

The logic of H_{ii} assumes that item answers should rank two people in the same way that total score ranks them—i.e., the person who passes an item should have a higher score than the person who fails it. H_{ii} is equal to the percentage of such correct discriminations minus the percentage of wrong discriminations. The relation between H_i and H_{ii} is not known.

In selecting the items for a test, it is proposed that all the H_{ii} values be calculated and that the items with low H_{ii} values be eliminated. This follows the usual procedure of item analysis. Loevinger then points out that if the heterogeneity (H_{ii}) is evenly distributed over all the items, 'we still cannot decide whether we have what Guttman calls a quasiscalable universe or two sub universes. To make such a decision we need a table showing the coefficients of H_{ij} . Apparently for a quasiscale these coefficients will be moderate in magnitude and fairly uniform. In case [among] the values of H_{ij} [there] are some very high and some very low despite uniform values of H_{ii} , we expect there is a way of dividing the items into two or more tests each of which is more homogeneous than the original test' (33, p. 520). Loevinger goes on to indicate that it is doubtful that items will often be easily separated into two or more homogeneous tests in this way, with H_{ii} 's all close to unity or zero. In other words, although the calculation of these coefficients is straightforward and more objective than some of the procedures which Guttman recommends for interpreting the coefficient of reproducibility, Loevinger is also faced with the difficulty of determining what shall be considered acceptable evidence of the existence of homogeneity or of a single dimension. She recognizes this fact and suggests that it must be determined in the light of the purposes at hand. A test which is sufficiently homogeneous for some purposes will not be good enough for others.

Loevinger has attempted to rule out the effects of difficulty on her coefficient of homogeneity by the device of dividing the value $p_{ij} - p_i$ by $1 - p_i$. This amounts to expressing $p_{ij} - p_i$ as a proportion of the maximum value it can attain at the given level of difficulty of the item. As a result, H_{ij} can vary from one to zero at any level of difficulty. This constitutes some gain over the am

biguities of the coefficient of reproducibility with respect to difficulty of items.

These coefficients, like Guttman's, will vary with populations, with situations, and with sets of items. Moreover, the same questions may be raised about the limits within which the scaling method is appropriate. To what extent will such scales be found for psychological processes that do not involve agreement with logically related statements or behavior? And is there any way to apply the same logic to unique individual processes not organized along the same dimensions in different persons?

Loevinger rejects Guttman's notion that scaling procedures discover anything about the scalability of a universe of which the test items are supposed to be a sample and insists that the investigator's judgment of what constitutes such a universe is not to be trusted; that scale analysis should rather be regarded as a method for the objective definition of psychological characteristics.

Finally, in a discussion of the relation of her techniques to factor analysis, Loevinger makes the point that a test showing high homogeneity cannot be expected to be factorially pure; that the criterion for a cumulative homogeneous test will be satisfied equally well by tests made up of items which measure a single factor or by tests composed of items which measure an approximately constantly weighted sum of factors. In other words, the dimensions singled out by this method are not irreducible, unanalyzable processes, but similarly patterned processes which act consistently in different people under certain conditions.

In summary, the scaling procedures discussed find evidence of functional unity in the consistency with which item responses order themselves. When respondent answers are the data and the tests are cumulative (Guttman, Loevinger), orderliness is seen in the fact that all people who agree with some item *A*, also agree with item *B*, but not all who agree with *B* agree with *A*. *A* is regarded as having a position which must be touched on the way to *B*, just as the first milestone must be passed enroute to the second. When judges agree on the position of statements with respect to some dimension such as favorableness (Thurstone scales), this suggests a common perception of a dimension on the part of judges.

Whether such organization has any relation to the functional unity implied by scaling of respondent answers is a problem for empirical test. The available evidence does not suggest close correspondence between scales constructed by these different methods.

Scaling methods provide a more systematic and rational method of studying the organization among items than do traditional item analyses. Another advantage is found in the fact that when items are shown to scale, a rational method of weighting the items is provided. As long as the weights assigned are equal or increase or decrease consistently with the scale position of the item, persons obtaining a given score will do so by answering the same items and a change in weights will not change the relative positions of scores.

The following important limitations of scaling methods must be kept in mind.

(1) The confusion between the effects of difficulty levels and of heterogeneity of process on indices of unidimensionality or homogeneity has not been wholly eliminated. The ambiguity which this introduces into the interpretation of the coefficient of reproducibility has been mentioned. Loewinger's attempt to correct for this effect yields an index varying from one to zero but the properties of the index between these values is not clear.

(2) It must be noted that in these, as in all other methods of investigating interrelations between items or tests, the effects of variable error cannot be separated from the effects of heterogeneity of processes actually involved in the reaction. This will be discussed in the section on reliability.

(3) The functional unity reflected in a set of scaled items is very evidently dependent on such conditions as the sample used and the structure of the testing situation. A scale which exists with one sample and under certain conditions may disappear with a new sample and altered conditions, a fact which recalls the statement that the standard meaning of any instrument must be defined with respect to a set of clearly specified determinants.

(4) Scaling alone tells nothing about the character or complexity of the processes which are responsible for the observed organization. The observed functional unity may be due to a complex of factors varying together under the specified conditions of the test or it may

stem from a simple, unitary process. The interpretation of the nature of the process underlying such functional unity will be discussed in the section on validity.

(5) When a series of items has been shown to scale by the Guttman or Loevinger method, a kind of functional unity is demonstrated to exist among the items. It is questioned, however, whether such methods will be successful in discovering the kind of process organization which is reflected phenotypically in statements that do not bear an obvious logical relation to each other. Moreover, the functional unity found by these methods can be demonstrated only when the same kind of organization is common to the whole population tested. Perfect scales are not obtained when there are unique patterns among those in the sample.

(6) Finally, it must be asked whether the analysis of complex psychological processes into their unidimensional components (Guttman) is necessarily the best and only method of observing and measuring them. It is probable that an inferred process such as favorableness or unfavorableness toward some object may be reflected in different ways in different people and that a comparable amount of favorableness or unfavorableness may have its source in different processes in these various persons. If this is true, the heterogeneous test (Likert, Thurstone) might best serve the purpose of discovering the strength and direction of the attitude complex as a determinant of behavior.

INTERTEST CORRELATIONS The literature contains many instances of the familiar technique of seeking for communality of process by correlating scores made in different situations or on different tests. To take a recent example, Adorno, Frenkel Brunswick, Levinson, and Sanford (2) report relatively high degrees of relationship between tests of various kinds of ethnocentrism (E scale), and between ethnocentrism and a test of fascist characteristics (F scale). May and Hartshorne employed the same procedure in their futile search for a general trait of honesty (26).^a

^a When we speak of intercorrelations as a method of demonstrating functional unity it is obvious that the many different correlational procedures rest on essentially the same logic. The special problems of methods such as Pearson's r 's chi square analysis of variance etc. are discussed in standard statistical works. It should also be noted that communality of process may be sought within the individual as well as between tests. See discussion in this chapter p. 276 and Cattell for other patterns of co-variation (8-9).

The "split half" and "alternate forms" methods of determining "reliability" become special instances of measuring functional unity by intertest correlations. A deliberate attempt is made, of course, to devise "comparable" stimuli or items, which presumably tap the "same" process, but the evidence that this has been accomplished rests on the correlations obtained. The same correlations become evidence of reliability. There is no satisfactory way of separating the effects of the differences in test items in two test forms from uncontrolled conditions which introduce random variation into the responses (unreliability). The isolation of the latter is the purpose of reliability tests, which are discussed in a later section of this chapter.

A number of familiar problems again emerge as limitations or as sources of ambiguity in interpreting these intertest correlations as evidence for the presence or absence of functional unity. Here again the effects of difficulty level (or frequency of agreement level) of the test items may affect the correlations and obscure, or enhance misleadingly, the true relationship stemming from functional similarities in the psychological process underlying test responses. For example, if one test is made up of items with which a large proportion of the sample population agrees and the other test has less generally accepted items, the correlation between the two tests will be reduced by this fact. The coefficient cannot equal unity even though the same processes are common to the two tests (6). But under other conditions misleadingly high correlations may result from a confusion of the effects of heterogeneous process and different difficulty levels. If a sample of children six to twelve years of age were given a reasoning test and a test of motor skills, each made up of items covering a range of difficulty, the positive correlation that would result would not imply any integration or interrelation of the processes of reasoning and motor skills but would be a function almost exclusively of the fact that tasks of greater difficulty fall within the competence of older children whether the tasks involve reasoning or motor response. This calls attention to the importance of knowledge about and intelligent selection of samples when interpreting the meaning of intertest correlations as indices of functional unity. It also suggests the desirability of equating instruments for difficulty before testing their intercorrelations.

The sample populations to which the tests are given may vary

in two ways—in the qualitative characteristics of persons sampled and in the distribution of scores on these characteristics—and the correlations may be affected by either difference. The California study (2) illustrates both these effects. Correlations between ethnocentrism and political economic conservatism ranged from 0.14 for San Quentin prisoners to 0.86 for a group of working class women. It is clear from the data available that correlations are higher for the populations having larger standard deviations, but there is also evidence that the size of the coefficients is related to the kinds of groups sampled. The authors suggest, for example, that the low correlation between ethnocentrism and political attitudes in the prisoners may be due to the inadequate frame of reference which these men have as a basis for evaluating political and economic events (2, p. 836). This means then, that evidence of communality of process does not provide conclusions generalizable beyond the samples used unless there is empirical evidence or a theoretically grounded basis for knowing what the conditions of generalization to other samples are.

Here, as elsewhere, the degree of relationship indicated will be a function of the specific statistical procedure selected for expressing the relationship. This selection should be made in the light of the characteristics of the data and the uses to be made of them. If the data are essentially qualitative in nature, the relationship between the tests may be expressed simply in terms of the proportion of individuals correctly classified in certain categories of test *B* on the basis of knowledge of their classification in test *A*.

The various types of correlation coefficients or regression equations will express the relatedness, if the data are quantitative and if the statistical procedures are appropriate. It is possible, as a rule, to determine whether a distribution of test scores conforms to the assumptions of normal distribution, linearity, and homoscedasticity which are implicit in many of the correlation techniques. But it should be noted that even though these assumptions are met for distributions of scores expressed in arbitrary units, this does not constitute proof that the variables reflected in these scores have the same characteristics. This is to say that an unknown amount of distortion is introduced into measurements which employ a mathematical model to deal with variables that do not have the characteristics of the model. If the test scores misrepresent the variables involved in a correlation, the coefficients of correlation based on these scores

may not accurately represent the relationship between the variables. But as will be shown later, the ultimate test of the validity of any construct and of the measures which enter into its definition is found in the utility of the construct in the process of reducing the matrix of events to some meaningful order. If constructs derived from correlations prove to have value as parts of a dynamic system, this suggests that the mathematical model is appropriate in some degree.

It has been indicated that a major problem in the construction of any instrument must be faced when combining responses to a number of items to produce a total score. The problem must be recalled briefly at this point because the particular solution selected in any given case will affect intertest correlations and thus the evidence of communality in the tests. The earlier comments on proposed solutions to this problem are relevant in the present context. Only one additional point need be elaborated here. It was suggested that most methods of assigning weights lack a theoretical rationale for determining the manner in which items of observed behavior should be combined to produce a score. The point is this: the satisfactory determination of weights to be given items in a test and the manner in which they are combined to produce a score should rest on knowledge of how the processes tapped by the items combine dynamically to produce effects. Coombs (13) has summarized concisely the way in which distinguishable processes may enter into combinations: (1) Processes may be additive, so that a greater quantity of any one will produce increased effects. (2) They may be conjunctive, that is, each must be present in some minimum amount before any effect is produced and no amount of one will compensate for the absence of the other. For example, both knowledge and motivation are necessary in solving a problem. (3) They may be disjunctive. One does not occur if the other occurs. This is best illustrated by the substitute mechanisms—*viz.*, hostility may be expressed by reaction formation, by direct aggression, or by displaced aggression. This means that some items may properly be added, but that others which are not additive must be combined in such a way that they represent the result which actually emerges. Until analyses of this kind are possible, the arbitrary scores obtained by combining test items have limited uses and the meaning of the correlations of such tests is ambiguous.

When all these problems have been taken into account, an

obtained correlation between two tests simply demonstrates that some functional unity exists, within the limits set by the size of the coefficient and by the ambiguities of its interpretation. From this information alone it is impossible to say what the common aspect is and whether it is simple or complex. Moreover, no method is provided for isolating and measuring the homogeneous process, whatever it is, for test scores do not rank persons with respect to the common factor unless the correlation approaches unity.

To summarize, correlation between tests has often been used as evidence of functional unity. The existence of correlation does not, however, illuminate the nature of this underlying process, its simplicity or complexity. Only when these indices approach unity (with allowance for unreliability) is it possible to rank persons in terms of the common process.

All such evidence of functional unity must be interpreted in the light of the limitations set by the following facts: (1) Correlations are influenced by the difficulty levels of the tests correlated. (2) Correlations are a function of the sample tested. (3) The statistics expressing covariations must be appropriate to the test data if they are to reflect the true relationship between the variables being measured. (4) The weights assigned to test items often affect correlations and should be determined by knowledge of the manner in which the psychological processes tapped by the items combine to determine a composite result.

FACTOR ANALYSIS It was inevitable that sooner or later the attempt should be made to reduce the accumulating mass of knowledge about intercorrelations between tests to a smaller and more manageable number of dimensions. The process of factor analysis involves a demonstration by mathematical procedures that if there were certain underlying processes or factors x , y , and z , a given table of test intercorrelations could be reproduced. It furnishes a statement of the number of factors needed to account for the matrix and the average loading of each factor on each test.

The assumptions in factor analysis which concern us here are those that enter into the interpretation of the results of factorial studies as means of discovering functionally unitary processes. Since the data of factor analysis are tables of correlation coefficients, the assumptions underlying those statistics are involved at the outset, and this is a hurdle which is largely ignored. There are, in addi-

tion, the following assumptions basic to methods such as those of Spearman, Thomson, and Thurstone. (1) When the factors contributing to a test score are uncorrelated, the correlation between the tests is equal to the sum of the products of the weights of those factors common to the tests. (2) The component processes combine by simple addition to produce the response made to the test. (3) Every person taking the test is assumed to possess every factor. (4) Finally, the weight of any factor on a test is the same for every person.

Whether the factors resulting from an analysis will be correlated depends on an arbitrary decision by the investigator to use or not to use orthogonal dimensions. If he does so, the statement made in the first assumption will follow from the nature of the mathematical relationships. The fact that it must be decided whether to regard the factors as correlated illustrates an important aspect of any of the attempts to discover functional unity. Although the data themselves set certain limits on what can be discovered, the hypotheses which the observer brings to the analysis will also determine what is found. Spearman and his group have argued that the factors should be uncorrelated, whereas Thurstone, among others, has insisted that factors which are the result of the complicated and interrelated processes of growth and experience should logically show some correlation. Since either type of factor may be found with the appropriate methods, the decision to seek one or the other must be made in terms of hypotheses which are independent of the process of factor analysis.

Although the second assumption—that component processes combine additively—apparently yields the same results as would a multiplicative assumption (38), a question may be raised regarding the adequacy of factorial methods for discovering all the kinds of organization in which we are interested. It may be plausible to assume that abilities summate, but it is not clear that this is a justifiable basis for predicting the combined effects of all variables. It is equally reasonable to assume relationships of conjunctivity or disjunctivity. In any case, a plausible theory is needed before simple additive assumptions can be justified.

It is difficult to obtain clear-cut evidence regarding the justifiability of the third and fourth assumptions—that every person taking a test possesses every factor, and that the test has the same loadings

on each factor for each person. When we are dealing with inter-individual variation it appears to be impossible to do more than estimate an individual's score on a given factor, and these estimates have unknown probable errors (44). In discussing this problem, Wolfe reports the observation that different people use quite diverse methods of attack on the same problem. This suggests that attributing the same factor weights (averages of individual weights) to all persons almost certainly produces distortion of the true situation in the individual case (44, p. 4). The newer *P* technique based on correlations of data from one individual's performance on the same tests administered on different occasions may reveal the comparative weights of given factors in different individuals, in so far as factors may be correctly identified as the *same* in these different individuals (9 pp. 27ff).

In addition to the problems raised by the assumptions underlying factor analysis, a number of other familiar questions must be faced. There is, first of all, the recurrent problem of distinguishing the effects of difficulty level and of process similarity or difference. Since factor analysis starts with a matrix of correlation coefficients there is the possibility at the outset that the size of the coefficients has been affected in the manner described above by the distribution of item difficulty in the correlated tests. Furthermore, it is clear that factorial methods may discover factors which are due largely to difficulty level (or popularity level). Ferguson has shown, for example, that when a matrix of intercorrelations between the items of a homogeneous test (by Loevinger's criterion) is factored, there are as many factors as there are levels of difficulty in the test (18). It would be expected that where tests in a matrix represent various levels of difficulty, common factors would be discovered in those tests at the same level of difficulty. This may constitute a means of separating difficulty factors from others, but it may also produce misleading conclusions.

Factorial methods, like others, have been shown to be affected by the samples used. The factors and the factor loadings discovered in one sample will often be different in a new population with different characteristics. This is not surprising, since "factors are produced by anything that introduces correlation into a set of variables" (44, p. 25). This means, then, that the relevant characteristics of a sample providing the data for the original factoring must

be known if we are to be able to specify the conditions under which the obtained factors exist

Another problem is found in the fact that any matrix may be factored into an infinite number of different components depending on the method used. In other words, no one solution is unique. The choice of any set of factors is, therefore, quite arbitrary and will be made in terms of the theories and hypotheses adopted by the investigator (8, pp 281ff). The factorial methods like others yield evidence of the existence of functional unity. Something is common to reactions in different situations. The lines of cleavage are found in different places by different methods of analysis, and the resultant slices of psychological reality are not always the same. The decision about which products are best, what they mean and whether the processes are simple or complex can be determined only in the light of some integrated theory of personality and behavior and not by statistical analyses alone.¹⁰

In summary, the principal purpose of factor analysis is to reduce a matrix of correlations to the smallest possible number of dimensions in the interest of parsimonious description of the interrelationships between the variables. The factors discovered may lead to fruitful hypotheses to be tested by experimental methods. For example, it might be demonstrated that a factor in certain tests emerges under ego involving instructions and disappears when the instructions are changed. Such studies would greatly increase the significance and scientific utility of the patterns of organization discovered by this method.

There are a number of sources of ambiguity in interpreting the results of factorial studies. (1) The appropriateness of certain assumptions underlying factorial methods may be questioned. (2)

¹⁰ Limitations of space as well as the practical difficulties in using the method in its present form have led us to omit a discussion of the special features of Lizardsfeld's promising technique of latent structure analysis (40). Its most important difference from the traditional methods of factor analysis lies in the fact that it is adapted for use with nonparametric data and so avoids some of the unwarranted assumptions about units of measurement implicit in the earlier methods. It is applicable to data collected by the method of single stimuli. Like other methods it demonstrates only the compatibility of the data with the existence of certain underlying trends. It does not isolate these for measurement or determine what they are. Coombs and his students are also engaged in devising methods appropriate to nonparametric data (5, 12).

The solutions obtained are not unique, the factors that are discovered are a function of the hypotheses of the investigator (3) The factors found may be due to anything which introduces correlation between variables, and this may be a common level of difficulty rather than a fundamental process of some kind (4) Factors identified are a function of the sample used and of the conditions of the observations (5) Factor analysis like the other methods, cannot solve the problem of isolating error variance from other sources of variation (6) Considerations other than the procedures of factor analysis must enter into the interpretation of the meaning of the factors discovered These will be discussed in the section on validity

DYNAMIC ORGANIZATIONS In the introductory remarks to the section on functional unity, it was suggested that the phrase has three possible meanings concomitant variation, dynamic interdependence and dependence of one process on the other (cause and effect relation) The techniques which have been described thus far all involve the demonstration of concomitant variation A question should be raised at this point about the limitations of such procedures for establishing all the types of functional organization which are of interest to social psychology The complex constructs of every science are defined in terms of dynamic and dependent relationships between identifiable segments or events Force is a set of relationships the atom is likewise a system of relationships An attitude is a complex of cognitive and affective processes seen as related to an object, or to an object image, and cohesiveness can be described only in terms of interrelationships in a group The actual dynamics of the psychological system are rarely known but are nonetheless implied

Two problems will be noted at this time (1) Although concomitant variation may be expected to appear where events are interdependent or where one is a determinant of the other, the fact of correlation is not in itself sufficient to establish the dynamic or dependent character of the relationships between variables (2) A distinction should be made between correlations based on observations of interindividual or intergroup variation as against intra individual or intragroup variation, for only in the latter case can events be shown to *change* interdependently

The distinction between concomitant variation, dynamic interdependence, and causal relationships is a familiar one The existence

of correlation merely indicates that two or more events change together. It cannot be inferred from correlational evidence alone that variables are interrelated dynamically in the way that every part of the surface of a bubble depends on every other part for the maintenance of the integrity of the whole, or that one event depends on the others in the sense that raising the temperature of a gas increases its volume. Projection, for example, is defined conceptually as a relationship between a person's perception of hostility in others and the presence of repressed hostility in the perceiver. Investigation would presumably reveal correlation between these two variables. This evidence, however, would fall short of providing a satisfactory demonstration that repressed hostility in the perceiver and perception of hostility in others are dynamically or dependently related and that there is in fact a process, projection, which can be defined by this relationship. In the absence of controls which would rule out common determinants of these processes and make possible their manipulation, their actual interdependence must be regarded as hypothetical.

Hovland and his collaborators (29 pp. 72-73) have discussed this problem, noting that the independent manipulation of two processes is often difficult, if not impossible. This is true especially when the variables involved are constructs that must be inferred from their effects. These authors were concerned with the question of determining whether soldiers' motivation to go overseas was affected by their attitudes toward the British. Specifically, the question is: if you show movies depicting the British role in the war and find that both motivation to go overseas and attitude toward the British change, is there any way to find out whether change in attitude has been responsible for change in motivation? Since correlation between variables is the only evidence, it is not clear whether motivation has been changed by the movies directly or by attitude change, nor is it clear whether attitude affects motivation or motivation affects attitude, or whether they interact. To establish such relationships, control of the independent variable is necessary, and in this case impossible. In the absence of such controls, we can only surmise about the nature of the relationships. The problem is ubiquitous and familiar, but often overlooked.

Another familiar problem must be mentioned. How far can we go toward establishing dynamic systems in the absence of an ade

quate metrics, one which is appropriate to both the data and the theory of social science? Devising an appropriate metrics becomes particularly difficult when the theory and data of a science make different demands. As already indicated, psychological data do not for the most part fit the assumptions of ratio scales and yet many theoretical concepts—particularly motivational ones, are based on models that assume processes differing in energy and amount which cancel or reinforce each other and which are dispelled in substitute activities. If we are to describe such constructs accurately and devise standardized instruments for their measurement, decisions will have to be made regarding the precise character of the relationship between the component processes whose relations define such complex constructs. This is illustrated in such concepts as displacement and other Freudian mechanisms, or the convergence and opposition of vectors in Lewinian systems or increments of habit strength, drive strength, and effective action potential in the Hull model. The problem is as yet unsolved, and it can be said only that the crude and indecisive tests of relations possible in the present state of measurement techniques do not in fact yield anything approaching a dynamics of behavior. We talk about vectors, forces and energy but we do not measure them.

When correlations are based on data derived from the observation of two or more processes in a number of individuals, the correlational evidence of organization of those processes reflects, in the very nature of the case, only the static relatedness of a cross sectional sample of events, such as one sees in a single frame of a motion picture. If procedures are to be developed for dealing with functional unities such as motivation and drive, attitudinal structure, substitution processes (displacement, sublimation), and group cohesiveness and attractiveness, attention must be given to more than a pattern of static relations such as are available in data on interindividual or intergroup variation. The parts of the pattern must be observed as they interact and change together. This means observing change in the same person or in the same group.

In other words, the attempt to manipulate variables through sampling individuals or groups is analogous to establishing Ohm's law by measuring resistance, potential, and current at one value level in one circuit and at a number of other value levels in that number of additional circuits. The alternative is, of course, to observe change

in the *same* circuit, or the same individual, or the same group. The two techniques should yield the same results if all the relevant factors that might affect the relation between variables are the same from circuit to circuit, or person to person, or group to group, but this is unlikely to be the case, particularly in the study of organisms or societies. This problem can be handled only by the study of organization and change within the individual, where, for example, the interdependent alteration in various aspects of an attitude may be observed, despite the fact that the conditions of the change are very different from person to person.

Many psychologists have, of course, been concerned with the observation of individual dynamics. Clinicians and other students of personality often use this approach exclusively, and more recently attempts have been made to apply factor analysis to observations of variation within the single individual.¹¹ Cattell suggests that data for this purpose may be obtained by observing responses of the same individual to the same tests under conditions which change either in accordance with some systematic plan or simply under the influence of the uncontrolled factors in the situation (*P* technique) (8, 9). Presumably any of the procedures discussed thus far could be adapted to such data to discover which of the reactions of an individual (to items, tests, etc.) show concomitant change, common factors, or unidimensionality, but the practical difficulties are great.

Similar problems arise in the attempt to discover functional unities among group phenomena, though very little systematic attention has been given to the matter. For example, group characteristics such as cooperativeness, low intragroup aggressiveness, the perception of uniform attitudes, the absence of cliques, and high output have been regarded as indices of group "cohesiveness." Sets of variables of this sort have been manipulated experimentally and observed to affect intragroup behavior (21), but there is little evidence that such variables change together, so that we are justified in identifying a functionally unitary construct to be called cohesiveness. If this were done, it would be analogous to studying organization within the individual. On the other hand, the correlation of characteristics between groups is comparable to looking for functional unity in the covariation of characteristics as you move from

¹¹ See the description of other possible patterns of variation (8, p. 96; 9, pp. 28ff.).

individual to individual. Thus, comparative studies of cultures, such as Murdock's correlation of kinship systems and incest taboos (34), take the relatively static approach of showing that, as cultures are now constituted, some characteristics go together. Here, again, it may be questioned whether a dynamics can be established from these sorts of data.

The Problem of Functional Unities

The problems which have been discussed in this section are those that arise when the attempt is made to discover processes which can in some measure be bounded and set apart from the on going stream of events for the purposes of scientific observation. Any common set of identifiable characteristics may be the basis of classifying and isolating events, but the unity discussed here is that discovered by techniques for demonstrating concomitant variation, interdependence or dependence of events on each other, item analysis scaling, interest correlation, factor analysis, and demonstration of dynamic relations.

There is no evidence from any of these methods of analysis that psychological unities are indivisible and ultimate or that we shall ever have such units. Moreover, the parts of psychological reality do not break out cleanly or stably, and the foci and character of organization change with the specific conditions under which responses are elicited—i.e., with the instructions given, the questions asked, the perceptions by persons in the sample of the meaning of the situation, their attitudes, needs, and capacities. In other words, the organization of psychological processes and the functional unities reflected in response must be thought of not as rigid, clearly delimited segments, like atoms or genes, but rather as shifting and somewhat unstable events, more analogous to wave patterns in a fluid medium or currents in a mass of air. The medium sets limits on the patterns that can arise, on their stability, and on the forces that can be generated, but a host of contingencies external to the medium also helps determine the character of the patterns and the magnitude of the resultant forces.

The patterns revealed also vary with the methods of determining functional unity, which do not necessarily yield the same segments of psychological reality. Responses to Thurstone scales often

break into multiple factors when factor analyzed (20), and more than one factor emerges from tests scaled by the Guttman technique (40, p 201) Test items chosen for their high indices of discrimination do not necessarily scale, and a correlation matrix may be broken many ways It is important to discover when different methods yield the same results, but the fact that they sometimes fail to do so does not necessarily discredit any procedure Other things being equal, the decision about which method is best must depend on the scientific fruitfulness of the constructs produced This problem will be discussed in the next section

Generally, the methods of discovering functional unities simply reveal the presence of concomitant variation Moreover, a large proportion of these investigations involves analyses of variation between individuals It is suggested that the identification of dynamic interrelations requires the observation of change within the individual or group rather than interindividual or intergroup variation But the statistics of correlation alone are not adequate to deal with problems of dynamic organization, even when applied to data from the single individual The definition of such dynamic constructs requires the demonstration of actual interdependence and interaction between part processes and a statement of the nature of these relationships Establishment of such relationships necessitates the experimental manipulation and control of the variables in question and a metrics that is appropriate to the data and to the theory of the science

The discussion has dealt largely with individual psychological processes because the problem of functional unity has been explored most extensively in this area But questions about functional unities, which are the fundamental constructs of our thinking must be answered for all areas of social psychology, including interactional processes group structure and related topics

THE INTERPRETATION OF FUNCTIONAL UNITIES VALIDITY

It is obviously not sufficient to stop with the discovery of functional unities The processes isolated must be interpreted and given meaning if they are to have use in a scientific system This is the

problem of validity. Although the distinction between functional unity and validity seems clear enough when so stated, it is not, in fact, precise. In both cases we are concerned with relationships between events, and many of the problems involved in the search for units are the same as those in establishing validity. Furthermore some of the procedures which we have identified as appropriate for the discovery of organized processes have often been regarded as tests of validity. For example, the demonstration of internal consistency by item test correlations has been called an index of item validity. Similarly, tests have been validated by correlation with other tests, although these methods have been discussed in this paper as evidence of functional unity. The notion of dynamic organization comes even closer to the complex of interrelations involved in validity studies. The distinction is then, not clear cut.

A case might thus be made for the unimportance of the distinction between functional unity and validity, but it is convenient for our purposes to separate two aspects of the question of relatedness of process, whatever terms are employed to identify them. One is the question of isolating parts with sufficient integrity for study and investigation. The other is the problem of giving these parts meaning by finding the way in which they fit into the whole pattern of events. The first question tends to push toward the discovery of simpler (though not necessarily irreducible) segments. The second raises issues about the general structure of the science. Ultimately the two aspects must be neatly integrated, for, as we have already suggested, the choice of part processes for investigation should be made in terms of overall theory, and the nature of integrated theory will depend in part on the components which are identified. Meanwhile, our point is simply this: when indices are used to indicate organized and more or less isolable process, they are regarded as tests of functional unity. When, however, a given item or test or cluster of tests or molar segment is the unit of organization, other tests or items or processes are considered in some sense discrete. To discover the complete network of relationships of any variable to other variables external to it is to give it meaning, to explore its validity.

Face Validity

It is possible to name and interpret any observed process or functional unity simply in terms of manifest similarities found in

the situations or responses which have been shown to be functionally related. This may be called face validity. It will be recalled, for example, that Guttman proposes that the universe measured by a scale be known by the content of the scaled items (40 pp. 53-54). This usually means that the referent of the attitude which is the common object of all the items is taken as the defining concept. If the investigator makes up a test consisting of items about Russia, the instrument, if it scales, is quite arbitrarily called a measure of attitude toward Russia. When May and Hartshorne (26) devised situations in which children could steal or lie or cheat, these concepts were defined denotatively. The behavior observed in the situation was what was meant by stealing, lying, and cheating. If the behaviors had been shown to be functionally unitary, they would have been interpreted by their common characteristic, deceit.

The interpretation of the factors discovered by factor analysis may be made in similar fashion. Thurstone, using the simple structure principle in order to arrive at psychologically meaningful factors, has this to say:

In order to interpret the primary factors it is usually necessary to examine the tests which have high saturation on a factor and to discover what is common to them. When such an hypothesis has been found it must be checked by examining all of the tests which have zero or nearly vanishing saturations on the factor. When a factor is fairly well understood then its presence or absence can be predicted with some confidence in a new test which has not hitherto been investigated factorially (12 p. 7).

Thus, if a reference axis runs through a group of tests manifestly involving verbal skills, this suggests that the common factor is a verbal factor.

There can be no objection to observing certain behaviors and stating that they are what we shall mean by honesty or extroversion or aggressiveness, provided we bear in mind the specific and limited meaning of the words. But certain cautions must be noted. More often than not, additional meanings are smuggled in and the assumption is made that the observations are, in fact, interpretable as a sample of a known universe—that, in other words, the instrument in question measures some process known to manifest itself in a specified universe of behaviors under a specified universe of

problem of validity. Although the distinction between functional unity and validity seems clear enough when so stated, it is not, in fact, precise. In both cases we are concerned with relationships between events, and many of the problems involved in the search for units are the same as those in establishing validity. Furthermore, some of the procedures which we have identified as appropriate for the discovery of organized processes have often been regarded as tests of validity. For example, the demonstration of internal consistency by item test correlations has been called an index of item validity. Similarly, tests have been validated by correlation with other tests although these methods have been discussed in this paper as evidence of functional unity. The notion of dynamic organization comes even closer to the complex of interrelations involved in validity studies. The distinction is, then, not clear cut.

A case might thus be made for the unimportance of the distinction between functional unity and validity, but it is convenient for our purposes to separate two aspects of the question of relatedness of process, whatever terms are employed to identify them. One is the question of isolating parts with sufficient integrity for study and investigation. The other is the problem of giving these parts meaning by finding the way in which they fit into the whole pattern of events. The first question tends to push toward the discovery of simpler (though not necessarily irreducible) segments. The second raises issues about the general structure of the science. Ultimately the two aspects must be neatly integrated, for, as we have already suggested, the choice of part processes for investigation should be made in terms of over all theory, and the nature of integrated theory will depend in part on the components which are identified. Meanwhile, our point is simply this: when indices are used to indicate organized and more or less isolable process, they are regarded as tests of functional unity. When, however, a given item or test or cluster of tests or molar segment is the unit of organization, other tests or items or processes are considered in some sense discrete. To discover the complete network of relationships of any variable to other variables external to it is to give it meaning, to explore its validity.

Face Validity

It is possible to name and interpret any observed process or functional unity simply in terms of manifest similarities found in

the situations or responses which have been shown to be functionally related. This may be called face validity. It will be recalled, for example, that Guttman proposes that the universe measured by a scale be known by the content of the scaled items (40 pp. 53-54). This usually means that the referent of the attitude which is the common object of all the items is taken as the defining concept. If the investigator makes up a test consisting of items about Russia, the instrument if it scales is quite arbitrarily called a measure of attitude toward Russia. When May and Hartshorne (26) devised situations in which children could steal or lie or cheat, these concepts were defined denotatively. The behavior observed in the situation was what was meant by stealing, lying, and cheating. If the behaviors had been shown to be functionally unitary, they would have been interpreted by their common characteristic: deceit.

The interpretation of the factors discovered by factor analysis may be made in similar fashion. Thurstone, using the simple structure principle in order to arrive at psychologically meaningful factors, has this to say:

In order to interpret the primary factors it is usually necessary to examine the tests which have high saturation on a factor and to discover what is common to them. When such an hypothesis has been found, it must be checked by examining all of the tests which have zero or nearly vanishing saturations on the factor. When a factor is fairly well understood, then its presence or absence can be predicted with some confidence in a new test which has not hitherto been investigated factorially. (42 p. 7)

Thus, if a reference axis runs through a group of tests manifestly involving verbal skills, this suggests that the common factor is a verbal factor.

There can be no objection to observing certain behaviors and stating that they are what we shall mean by honesty or extroversion or aggressiveness, provided we bear in mind the specific and limited meaning of the words. But certain cautions must be noted. More often than not, additional meanings are smuggled in and the assumption is made that the observations are in fact interpretable as a sample of a known universe—that, in other words, the instrument in question measures some process known to manifest itself in a specified universe of behaviors under a specified universe of

conditions. Thus a test is assumed to sample adequately and in a systematic way the manifestations of some process called attitude toward Russia, or aggression or honesty, or cohesiveness. However reasonable the hypothesis may appear, the character and limits of the behavior universe sampled cannot be set merely by identifying a common characteristic of the items of a test and assigning to the measure the generality possessed by the name of that characteristic. The limits must be determined.

Furthermore, it is necessary to be certain that the manifest characteristic is the significant one. Since unobserved and uncontrolled factors may be responsible for the order among the items even in a test which scales such factors must be ruled out before it can be assumed what the common aspect is. It was suggested, for example, that perception of the logical relations between the items of the Guttman scale on Attitude Toward the Army may have been responsible for the appearance of a scale among the items, since people want to be logical. If this was true, it is misleading to consider that the order lies in the attitude itself. The same problems are present in interpreting factor loadings on tests. If the tests clustering around an axis are verbal tests, the fact that words are involved may well be the crucial fact. But other possibilities should be entertained and checked. These tests might, for example, have a common difficulty level which was different from those of other tests or they might be less fatiguing or more fatiguing than other tests and so on. One advantage of searching for the same factors in many different sets of data and in tests of the same kind which vary in minor ways is that accidental bases of correlation may be ruled out.

Prediction to a Criterion

The most familiar procedure for determining validity of an instrument is the simple demonstration of correlation between the measures made by the instrument and some criterion. When the correlation is high, the measure is said to be valid for *prediction to the stated criterion* and the test may be named by the criterion as in college aptitude tests or tests of attitude toward going to college.

The reader will recall the discussion of the factors which affect

the evidence of functional unity between two tests—such factors as the weights assigned to test items, the extent to which the variables conform to the assumptions implicit in the statistics used, the difficulty level of the tests in conjunction with the distribution of the measured characteristic in the population sample, the selection of specific mathematical procedures for expressing the relationship. All these problems are equally serious in validating a test against a criterion, and the reader is referred to the technical sources for more detailed discussion of them (4, pp 1245ff). Only two further points will be considered here: the problem of the selection of a criterion and the limitations of this method of validation.

The choice of a criterion will be determined by some practical or theoretical consideration. The practical need to select successful salesmen will dictate that some measure of success in salesmanship be taken as a criterion, or the demand for information on how many soldiers will go to college after leaving the Army quite obviously requires that a test of attitude toward this topic be validated against the criterion of actually going to college. Again, if the theory is held that lack of self confidence makes people poor judges of others, a test of self confidence might be validated against some measure of accuracy in judging others. In reality, the selection of any criterion is based on a theory about the relation of the measured process to the criterion situation or behavior, but typically the theory involved is some common sense assumption, such as the notion that a person who dislikes Negroes will be more likely to vote for segregating them than a person who likes them, or the assumption that the man who has a favorable attitude toward the church will be more likely to go to church than the one who expresses bitter resentment against the church.

It will be suggested that this level of theorizing is inadequate for the validation of tools or constructs of maximum scientific usefulness. For the present it need be said only that in the absence of an integrated theory of behavior which dictates the selection of validating criteria, literally any behavioral event or resultant may be chosen if it is itself to be predicted or if it is judged to reflect some process of interest. Practical common sense has often been the guide.

Any demonstrated relationship between a criterion and the data yielded by an instrument provides an additional fragment

of meaning, and, as the measures are shown to be related to still other criteria, they take on further significance. If, for example, a scale purporting to measure attitude toward Britain proves to be correlated with amount of historical knowledge, with church membership, and with income, we know more about the scale than we would if only one of these relationships had been established. Such information may lead to suggestive hypotheses about the source of the correlations and thus to further studies. A protest must be entered, however, against the proliferation of blindly empirical validities which are without the disciplined guidance of theory, for the increment of meaning from the accumulation of miscellaneous correlations may ultimately approach zero. The short run practical uses of this sort of validation procedure are evident in the designing of aptitude tests for selection purposes, but they do not appear to have provided the kind of interpretation needed in building a scientific structure.

This limitation is especially clear in attempts to validate measures of those important psychological constructs which are known as intervening variables (attitudes, motives, drives). It is useful to know that a questionnaire on attitude toward religion is answered differently by those who are church members and those who are not, but before such a concept can have any systematic significance, other steps are necessary. A theory about the structure and content of the attitude process and its interrelations with other processes in the determination of behavior must be worked out, and studies must be made to discover whether the hypotheses are supported. We turn then, to a discussion of this concept of validation.

Validations by Testing Predictions from Theory

The essence of the approach to validation through testing predictions from theory may be stated briefly. The meaning of any measured process is given not only by a description of operations used in isolating it from other processes and in assigning some index of quantity but also by knowledge of its influence on other processes and their influence on it. Consequently, to establish the validity of a construct and of the defining measures is to conduct experimental investigations. This involves all the problems of formulating theory,

deducing consequences, and testing the deductions under conditions of controlled observation.

If behavior theory leads to deductions about conditions of change in process *A* and the effects of *A* on other processes, ways must be found to determine the accuracy of these deductions. When predictions prove to be correct, both the theory and the construct as measured are validated in some degree. Where the predictions are seriously in error, it is difficult to spot the trouble unless either the theory or the method is sufficiently well established to be above suspicion. Thus, if the orbits of a number of stars and the theory of gravitation all point to the existence of an undiscovered stellar body, the failure to find the star will be attributed to imperfections in the instrument unless the investigator suspects that there might be conditions under which even widely accepted principles no longer tell the whole story. If, on the other hand, both measures and theories are suspect, untangling the source of the trouble will involve the clumsy methods of trial and error which are familiar to the psychologist. In any case, it is clear that validation of theory and of instruments of observation tend to proceed simultaneously and that they can be separated only in so far as experience has accumulated to suggest that predictions made from a given theoretical structure tend to work out well when the events involved are measured by one set of instruments and badly with another set or, conversely, that although a given method seems adequate in testing predictions from theories *A*, *B*, *C*, and *D*, things go wrong when predictions are made from theory *X*.

There are still other complications which make the interpretation of the unconfirmed prediction ambiguous. Just as the behavior and position of a star is a prediction from the interaction of many stellar bodies, so must behavior of organisms be predicted from more than one variable. This means that the measurement of any or all of the variables and the theory that relates each of them to behavior may be in error. In so complex a situation, it is no wonder that validation procedures began in the oversimplified and naive attempt to predict single criteria from single variables, in the hope that by luck some clear cut relationship would emerge.

Even though the confirmation of a prediction chalks up a score in favor of the theories and methods involved in the prediction, the

evidence from one such confirmation is of course, never conclusive. It must be supported by further checks in situations where the process measured in a certain fashion is combined with other variables to produce predictions of other consequences. By this procedure, the limitations of a construct, as defined by a certain method of observation, will become apparent as evidence accumulates regarding its successes and failures.

It is difficult to find examples among psychological studies which illustrate these points, because this method of validating measures has rarely been used deliberately, and never extensively. Certain aspects of the California studies of the Authoritarian Personality may serve, however, to clarify the problem. (2) The study was conceived around the theory that such psychological processes as attitudes toward outgroups toward authority figures toward discipline, and toward conventional morals are not independent but that the nature of these attitudes in a person is a function of some characteristic mode of adjusting to conflict and hostility, varying from relatively insightful, direct, and ego integrated attack to extremes of repression, projectivity, displacement, and ego alien mechanisms. It is predicted that the latter pattern of adjustment will produce ethnocentrism, conventionality, and docility before authority and that the first will underlie attitudes of tolerance for outgroups, acceptance of the unconventional and a more objective evaluation of authority. The evidence presented involves correlations between the phenotypic measures of these various processes. The correlations are roughly those that would be expected if the theory is true and the measurement relatively valid. However, it is obvious that this study does not provide a wholly satisfactory test either of the hypotheses or of the measures involved since there is no attempt to test predictions under conditions that systematically control the variables. The limitations are the limitations of correlational techniques in the study of dynamic organization, which were discussed on pages 278-282.

Stevens (39, pp. 47-48) has commented on the present predicament in validating measures of intervening variables, suggesting that we are now dealing largely with what he calls 'indicators,' related by unknown laws to the psychological dimensions in which we are really interested. We observe restlessness and infer drive, or we observe verbal statements and infer attitude. He goes on to say,

The difference, then, between an indicant and a measure is just this: the indicant is a presumed effect or correlate bearing an unknown (but usually monotonic) relation to some underlying phenomenon, whereas a measure is a scaled value of the phenomenon itself. Indicants have the advantage of convenience. Measures have the advantage of validity.

This may seem to suggest a somewhat more clear cut definition of validity than has been proposed here, by restricting it to the relationship between the measurement and the process measured. This is what we would really like to know, and it would be convenient if ways could be found to discover this relation without recourse to the complicated procedures of predicting from the measures and testing the predictions. The trouble is that there is no direct access to the underlying phenomena. It appears that we shall always observe indicants, for we cannot get inside and watch the attitude at work. The hope is that we shall approximate more and more closely the law which relates indicant and the thing we want to measure. That we have done so can be known only from the observation that if we assume some specific relation of process and measure, our predictions to other events are more accurate than when some other relation is postulated. The relationship must be guessed at and tested by its fruits.

We return briefly to a point which was mentioned in the early pages of this chapter. When evidence of the validity of an instrument has been obtained, it must be remembered that this validity has been demonstrated under a circumscribed set of conditions. These conditions may be explicit or implicit, their range may or may not be carefully spelled out, but they are, in any case, a necessary part of the description of the validity (or functional unity or reliability) of any instrument, for it is unlikely that any relationship is ever completely and unexceptionally true. There are limits within which it holds, and evidence should be provided about the conditions which must be constant and those that may vary without disrupting the prediction on which the evidence for validity rests. For example, it would be questioned whether the findings of the California study could be reproduced in some other culture, whether they hold for all motivational conditions, for all levels of intelligence. If they do not hold up, the scales are inappropriate for these

different populations or the relationships between the processes measured are different under the new conditions. The discovery of methods of isolating and controlling the relevant variables which are responsible for correlations and good predictions is the basic methodological problem.

STABILITY OF MEASURES RELIABILITY

The familiar concept of reliability does not require extended elaboration, and again the reader is referred to standard texts for a discussion of the technical problems (24-25). Two questions will be considered briefly at this point: (1) What is the meaning of reliability and how is it distinguished from functional unity and validity? (2) What are the limitations and the uses of the concept of reliability?

The notion of functional unity has been used in this discussion to refer to a functionally organized process, a kind of unit of observation and analysis. Validity adds the notion that such units, when properly identified, measured, and combined, will yield successful predictions to psychological events. Reliability raises the further important question of how stably such units or processes can be observed, measured, and inferred. When investigators inquire into the reliability of their observations, they are asking themselves how well they can control the determinants of the key response on repeated attempts to re-establish the same observational situation. It is a basic scientific assumption that when conditions are constant, the results must be the same. Consequently, if the results change, there must have been some change in conditions. When some response or inferred construct is consistently predictable from a specified set of conditions, reliability is high; when predictions are unstable, reliability is low.

The sources of unreliability will, of course, vary from situation to situation. Loevinger (32) has suggested two types of error which she calls content factors and transitory variations in efficiency. The content factors are found in the items which elicit response—i.e., in the instigating conditions. Loevinger actually interprets this source of error in a way that makes it a source of heterogeneity rather than of unreliability, in our meaning of the term, for she locates the problem in the fact that some of the test items involve

different abilities from those required by the test as a whole. But when the meaning of unreliability is taken as instability of reaction, the character of the items introduces unreliability in so far as they are not sufficiently specific in their effects to tap the same process or instigate the same reaction on different occasions. Loevinger's second type of error refers to the numerous uncontrolled processes in the individual that influence reliability: motivations, sets, fatigue, boredom, and similar variables which change from one testing to another.

Two other types of variable should be added which may modify the measured results from one observation to the next: (1) all the varying social and physical stimuli which affect the reaction, and (2) variations in the recording and interpretation of the behavioral events. This latter source of variable error may be isolated from the others by the familiar procedure of ascertaining coder or observer reliability through independent observations and independent coding of the protocols by several persons. The elimination of the other causes of variation that lie in items, in the total situation, and in the individual requires the application of the usual techniques of experimental control.

This leads us to underline an important point which is not new but which is usually ignored in practice (22, 12). It is misleading to speak of the reliability of a test or a tool with the implication that the reliability or unreliability is a property only of the instrument itself, for the error observed is the result of variation in the whole complex of determinants of the measured event. A procedure which produces stable measures under one set of conditions, or with one sample, or with one person may not do so with other conditions, with other samples, or with another individual. As indicated at the outset, any statement about what a test measures and how reliably the measurement is made must be accompanied by information regarding the *conditions under which the statement is true*.

It is necessary to examine briefly the operations for determining reliability in order to discover their relation to our conceptualization of the term and the attendant limitations of the procedures. The methods which involve correlation—the familiar test-retest, the odd-even, or split-half and alternate-forms reliability—indicate the stability with which individuals maintain their positions with respect to one another on repetition of the same tests or on different

sets of items which are assumed to measure the same process¹² The reliability of means, of differences and of other statistical indices attempts to estimate, in accordance with the theory of error, the probability that the true value of the index (*viz.*, a mean obtained from an infinite number of measures of the same universe) will fall within a certain range of values As Loevinger points out,

No matter how the reliability coefficient is computed the statistics of reliability assume that (a) the variable error factor has an expected or average value of zero (b) the error factor in one set of obtained scores is uncorrelated with that in another set however similar the test may be (c) the error factor in a set of scores is uncorrelated with the true scores and (d) the variances of the error factors in two comparable tests are equal (32 p. 6)

These assumptions obviously cannot be met if determinants of the reactions are changed in some systematic fashion from one observation to the next, as by changing test items, if residues of the first measurement affect the second, as when answers are remembered or sets changed in one direction by the initial testing, or if any other systematic nonrandom shifts in conditions occur from test to test And if the conditions cannot be fulfilled, then reliability indices give an inaccurate and ambiguous indication of the extent to which measures may be expected to fluctuate under the given condition

It is clear that these conditions are met by none of the operations for estimating reliability The test retest method which attempts to follow the logic of replicating observations under the same conditions (same items, same persons), runs into the difficulty that the measures on successive occasions are not independent, and that the variation from test to test is not wholly random in character On the other hand, alternate forms or split half procedures pose the problem of how to separate the effects of heterogeneity of process from unreliability When two or more forms of a test are being devised the attempt is made to construct parallel items that seem to involve the same process, and then their correlation is often assumed to be an index of reliability As Loevinger has insisted, however this solution is unsatisfactory It is necessary to provide

¹² See Loevinger (32) for a discussion of the logic of the methods of Kelley, Thurstone, Spearman, Brown, Kuder and Richardson and others

more objective evidence than this that the items do, in fact measure the same process. But if intercorrelation between the forms is taken as an index of reliability, it cannot also be regarded as indicative of the communality of process. Of course, where the correlations are near unity, it can be said that the instruments are measuring the same process or the same set of processes and also that they are measuring them reliably. But if the correlation is low, it cannot be clear whether this is because of the instability of the measures or because the different sets of items are in fact measuring different things. Our present techniques do not make possible the separation of these contributions to variance.

To sum up, then, it has been suggested that reliability be thought of as referring to the amount of stability which measures or observation reveal when repeated under conditions which ensure that only random variable errors affect this stability. To say that a measure or observation is reliable does not necessarily indicate that a significant variable is being measured, or one that we wish to measure or one that is uncontaminated by irrelevant influences. These are the problems of validity and homogeneity. To say that a measure is reliable means simply that the important determinants of the measured event—the instigating stimuli, the variables in the reacting individual, observational techniques and procedures for handling the observations and reducing them to the final result—are all sufficiently under control for us to be able to reproduce results within stated limits.

Whether an operation can be found which accurately represents the degree of stability of measures under conditions subject only to variable error turns on the problem of devising ways of replicating observations so that they will be independent of one another. Meanwhile in spite of the ambiguity of the indices which are used for this purpose, the justification for their continued use is this: even though we may be unable to estimate accurately the variable error on a test we can say that low correlations between a test and a retest indicate that conditions are not sufficiently stable to justify using the particular method until the causes of the instability can be identified and controlled. On the other hand, if the correlation is high the situation is a stable one that warrants further exploration in order to find out whether the source of the stability lies in some artifact such as mere recall and repetition of a previous re-

sponse, or in having found a way to control the determinants of an important variable which is the object of study

CONCLUSION

This chapter has reviewed certain problems of objective observation and a number of methods for dealing with these problems, with especial reference to their logic and their limitations. Attention has been given particularly to the concepts of functional unity, validity, and reliability.

It has been argued that all these problems must be seen and evaluated in the broad context of the assumptions and methods of science. Science busies itself with building comprehensive hypotheses about the relationships between events, deducing what must follow if these relationships are true and devising methods of observing the predicted consequences under conditions of controlled observation. To delimit and define constructs (functional unity), to interpret their meaning (validity), and to produce evidence of their stability (reliability) involves working within the framework of this logic. The design of objective instruments and procedures requires, therefore, a theory about the characteristics and relationships of any variable to be measured—i.e., its determinants, its dynamic interdependencies, and its consequents. The evidences of functional unity, validity, and reliability thus obtained are, like all scientific evidence, subject to the limitations imposed by the conditions of the observations, for the discovered characteristics of the observational procedures are contingent on those conditions.

BIBLIOGRAPHY

- 1 Adkins D C A rational comparison of item selection techniques
Psychol Bull, 1938 35, 655
- 2 Adorno T W *et al* *The authoritarian personality* New York
Harper 1950

- 3 Allport G W, and Odbert, H S Trait names a psycho lexical study
Psychol Monogr, 1936 47, No 211
- 4 Bechtoldt, H P Selection In Stevens S S (ed) *Handbook of experimental psychology* New York Wiley 1951
- 5 Bennett J F *Applications of isotagic geometry to the data of social science* Unpublished Ph D thesis Univ of Michigan 1950
- 6 Carroll, J B The effect of difficulty and chance success on correlations between items or between tests *Psychometrika*, 1945 10, 1 19
- 7 Carter, H Recent American studies in attitudes toward war *Amer Sociol Rev*, 1945 10, 343 352
- 8 Cattell, R B *Description and measurement of personality* Yonkers World Book, 1946
- 9 ——— *Personality, a systematic theoretical and factual study* New York, McGraw Hill 1950
- 10 Coombs, C H Some hypotheses for the analysis of qualitative variables *Psychol Rev*, 1948, 55, 167 174
- 11 ——— Psychological scaling without a unit of measurement *Psychol Rev*, 1950 57, 145 158
- 12 ——— The concepts of reliability and homogeneity *Educ Psychol Meas*, 1950, 10, 43 56
- 13 ——— Mathematical models in psychological scaling *P Amer Stat Assoc*, 1951, 46, 480 489
- 14 Dudycha, G J A critical examination of the measurement of attitude toward war *J Soc Psychol*, 1943, 18, 383 392
- 15 Edwards A L A critique of "neutral items in attitude scales constructed by the method of equal appearing intervals" *Psychol Rev*, 1946, 53, 159 169
- 16 ———, and Kilpatrick F P A technique for the construction of attitude scales *J App Psychol*, 1948 32, 374 384
- 17 Farnsworth, P R Shifts in the values of opinion items *J Psychol*, 1943, 16, 125 128
- 18 Ferguson, G A The factorial interpretation of test difficulty *Psychometrika*, 1941, 6, 323 329
- 19 Ferguson, L W The influence of individual attitudes on construction of an attitude scale *J Soc Psychol*, 1935 6, 115 117
- 20 ——— An item analysis of Peterson's 'War Scale' *Psychol Bull*, 1938 35, 521
- 21 Festinger, L., et al *Theory and experiment in social communication* Ann Arbor Edwards 1950

- 22 Goodenough F L A critical note on the use of the term reliability in mental measurement *J Educ Psychol*, 1936 27, 173 178
- 23 Guilford J P *Psychometric methods* New York McGraw Hill 1936
- 24 ——— *Fundamental statistics in psychology and education*, 2d Ed New York McGraw Hill 1950
- 25 Gulliksen H *Theory of mental tests* New York Wiley, 1950
- 26 Hartshorne, H and May, M *Studies in deceit* New York Macmillan 1928
- 27 Hinckley E D The influence of individual opinion on construction of an attitude scale *J Soc Psychol*, 1932, 3, 283 296
- 28 Horst, P (ed) The prediction of personal adjustment *Soc Sci Res Counc Bull*, 1941 No 48, 1 156
- 29 Hovland, C I Lumsdaine, A A, and Sheffield F D *Experiments on mass communication* Princeton Princeton Univ Press 1949
- 30 ———, and Sherif, M Judgmental phenomena and scales of attitude measurement I *J Abnorm Soc Psychol*, 1952, 47, 822 832
- 31 Likert, R A technique for the measurement of attitudes *Arch of Psychol*, 1932 No 140
- 32 Loevinger, J A systematic approach to the construction and evaluation of tests of ability *Psychol Monogr*, 1947, 61, No 285
- 33 ——— The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis *Psychol Bull*, 1948 45, 507 529
- 34 Murdock G P *Social structure* New York Macmillan 1949
- 35 Murphy, G. and Likert, R *Public opinion and the individual* New York Harper, 1938
- 36 Pintner, R. and Forlano G The influence of attitude upon scaling of attitude items *J Soc Psychol*, 1937, 8, 39 45
- 37 Sherif, M, and Hovland, C I Judgmental phenomena and scales of attitude measurement II *J Abnorm Soc Psychol*, 1953, 48, 135 141
- 38 Spearman C Abilities as sums of factors, or as their products *J Educ Psychol*, 1937, 28, 629 631
- 39 Stevens S S Mathematics measurement, and psychophysics In Stevens S S (ed) *Handbook of experimental psychology* New York Wiley, 1951 pp 1 49
- 40 Stouffer S Guttman L., Suchman, E A, Lazarsfeld, P F, Star, S A. and Clausen, J A *Measurement and prediction* Princeton Princeton Univ Press, 1950
41. Thurstone, L. L. Attitudes can be measured *Amer J Sociol*, 1928, 33, 529 554

42. ———. *Factor analysis as a scientific method*. Chicago The Psychometric Laboratory, Univ. of Chicago, 1951, No 65
43. ———, and Chave, E. J. *The measurement of attitude*. Chicago Univ. of Chicago Press, 1929.
44. Woffle, D. Factor analysis to 1940 *Psychometr. Monogr.*, 1940, No 3.

The Use of Documents, Records, Census Materials, and Indices

Robert C. Angell and Ronald Freedman

The research methods treated in this chapter differ from those treated in other chapters in that documents records census materials and indices characteristically bring data to the social scientist in a form over which he has relatively little control. In contrast when the social psychologist uses the method of observation either participant or nonparticipant he can focus on those aspects of the behavior of the subject population that have theoretical interest for him. If he uses tests or questionnaires he can choose or frame the instrument to suit his scientific needs. In interviewing the subject can be guided by the interviewer and if crucial points are unclear the interviewer can probe until the matter is elucidated. Although there are exceptions documents records, and indices normally come to the investigator ready made some other person either a participant in a social situation or process the originator of a system of recording or the creator of an index determined the

form of the data. These materials frequently have to be recast in terms of the research problem in hand before they are fully usable by the scientist.

There is perhaps another feature of these research methods that distinguishes them. Because the data come ready made, they do not depend upon the reach of a specific investigator or research team, whereas data that are obtained through scientific observation, through tests and questionnaires, and through interviews are gathered for a specific purpose characteristic of a particular research design and are drawn only from universes in space and time into which the formulators of that design can send investigators. Documents, records, and indices, on the other hand, may bring together data for scientific analysis from remote times and places.

DOCUMENTS

Since we are here concerned with research techniques pertinent to social psychology, the documents that will be discussed are those that give insight into processes of interaction. A document may describe a process of personal or group development; the only limitation on the complexity of the situations dealt with is that its writer must be able to embrace the situations adequately in his thought and treatment. The scientist must find in the account given the facts he needs to perform a theoretically satisfying analysis.

The category of documents thus delineated may be termed expressive documents. They are at one end of a continuum at the other end of which are such documents as court records, official histories and proceedings of commissions. In between are such types as newspaper stories and memoirs of diplomats which may give humanistic data but which rarely yield sufficiently detailed and embracing statements of the interactive processes. Expressive documents have been discussed at length in three research bulletins of the Social Science Research Council (3, 7, 26). Two of the monographs contained in these bulletins give extensive bibliographies (3, pp. 192-201, 26, pp. 164-73).

Types of Documents

We shall divide expressive documents into (a) personal letters (b) life histories either diaries autobiographies or what Abel (2) calls biograms (see p 304) and (c) accounts of small group process. As will appear in the discussions of the separate categories expressive documents are not always spontaneous for their production may sometimes be stimulated by the scientific investigator. This does not mean however that he necessarily imposes his definition of the situation upon the writer. If the investigator really wants a free and untrammelled treatment of the subject he will refrain from structuring it except for indicating clearly the area of life experience he wants covered.

Letters

Personal letters constitute the most frequently available type of expressive documents. Thomas and Znaniecki leaned upon them more than on any other source in their monumental *The Polish Peasant in Europe and America* (52). Series of letters were collected both by Thomas in the United States and by Znaniecki in Poland. 50 series 764 letters in all are reproduced in the text. They are used mainly to throw light on the changes in Polish peasant primary groups wrought by industrialization proletarianization of the peasants and emigration of family members. Although Blumer in his critique of this work states that the letters supplement the authors previous knowledge of Polish society rather than the other way round he concedes that the authors mulled over the letters a great deal and derived much from them in the way of ideas suggestions and generalizations which they incorporated into their theoretical statements (7 p 38). No other study has used this type of data so extensively.

The value of letters as expressive documents will vary with the culture in which the writers are immersed. It is probably true for example that personal correspondence was much more revelatory of life situations in pioneer America than it is today. Then there was neither the competition of recreational interests nor the opportunities to communicate by other means that have so largely robbed us of the art of letter writing. Only the separations of war have

in our day, revived somewhat the impulse to communicate one's inner life freely and fully in personal letters.

Life Histories

Diaries are a second type of expressive documents. Although they have been used extensively by historians and have been recommended by an eminent psychologist as "the personal document par excellence" (3, p. 95), they have not been used much in social-psychological research. Two examples in which attempts are made to formulate hypotheses, in part at least from diaries, are Runner's study of social distance in adolescence (40) and Cavan's study of suicide (11). There seems to have been no investigation that has drawn upon large numbers of diaries. Actually it would be very difficult to collect enough of them at the adult ages to investigate nomothetically a social-psychological problem, *i.e.*, to be able to develop scientific generalizations. Almost any scientific hypothesis would require the imposition of controls such as age, sex, and social class, and this would greatly limit the universe from which one could obtain the diaries. It is conceivable that one might stimulate the keeping of diaries by the group it was desired to investigate, but there would be a strong tendency for interest to vary so markedly that the documents would not be scientifically comparable.

Allport, in *The Use of Personal Documents in Psychological Science* (3), argues for the idiographic use of personal documents, including diaries, in scientific work. He points out that better predictions of the behavior of specific persons can frequently be made by careful study of their past than by applying to them generalizations drawn from the study of populations of similar people. However useful such idiographic use of documents may be for the psychological therapist, it would impose an unbearable burden on social psychology if it were adopted as the general route to scientific knowledge. It may be worth while to develop idiographic "laws" of behavior not only for maladjusted persons but also for a few outstanding leaders, such as Stalin or Nehru or Peron. The general advance, however, must come from scientific generalizations that are applicable to whole categories of people under specific conditions.

If diaries are not kept frequently enough to give much promise

of scientific usefulness, the case is even more discouraging with spontaneous autobiographies, and really perceptive ones from a social-psychological point of view are rare indeed. H. G. Wells wrote one of the few (62). Thomas and Znaniecki had real success in commissioning a life history by a Polish immigrant to America (52, Vol. II, pp. 1915-2226). This document goes into minute detail about almost every aspect of the writer's life. Clifford R. Shaw has likewise obtained excellent life histories from delinquent boys (43, 44, 45).

We should also include in the present category documents that are obtained by a person's telling his autobiography. This is not to be confused with "active interview" records, since in such an interview the respondent is stimulated to deal with those aspects of his life that are theoretically significant for the investigator. Cultural anthropologists have made the greatest use of spoken autobiographies. Kluckhohn, in his analysis of this technique, characterizes it as the "passive interview" (26, p. 125). He states that it is more appropriate to inquiries emphasizing personality, whereas the active interview is better suited to studies of the culture as such. Notable spoken autobiographies are one of a Navaho recorded by Dyk (20) and one of a Kwakiutl by Ford (21).

The interest in life-history documents has hardly ever been nomothetic in the strict sense. The investigator has wanted either to gather data that would enable him to understand the individual life—Dollard's criteria are designed for this purpose (18)—or to become generally acquainted with a culture or subculture. The Winnebago autobiography obtained by Radin (37) and the autobiography of a professional thief obtained by Sutherland (16) are good examples of the latter interest. It seems unlikely that in the future full autobiographies can be obtained from enough people to control variables and really test hypotheses. Such documents will probably continue to be of use in nomothetic research mainly by giving rise to "hunches" and very general hypotheses.

Another approach is furnished by Abel's prescription for collecting what he calls biograms (2). This is a term he coined to cover life records written, at the instigation of the social scientist, by a large number of persons who have been involved in similar experiences. Chalasinski used this technique in preparing his *The Young Peasant Generation in Poland* (13), and Abel himself used it for his *Why Hitler Came Into Power* (1). In the former case the research problem was to determine the role and significance of

youth organizations in peasant life, and in the latter it was to dis cover the factors leading to participation in the National Socialist movement. In both cases the documents were obtained by setting up a prize contest which offered a large number of awards. Anonymity was assured. Those who were eligible and the area of life to be covered were carefully specified in the announcements. Abel believes that this method provides freely written mass data on specific types of experience suitable for analysis by the social scientist. A somewhat similar technique was used in a study of German refugees from Hitler (4).

Accounts of Small group Process

Accounts of small group process by a participant are so rarely written spontaneously that they have not been used as the basis of any large investigation. In one case, at least, such documents have been stimulated by the researcher for a nomothetic type study. Angell obtained from University students 50 documents on family life before and after the impact of the depression (5). The writers were paid a small fee. They wrote from a rather broad outline which suggested aspects of family life that were to be covered. A sample topic was "Discuss the external conditions of your family's existence prior to the decrease in income. Touch on the type of neighborhood, the house and yard, the family's material possessions, etc." A number of hypotheses about family organization under stress were tested by analysis of the data.

Use of Documents

The peculiar value of expressive documents is that in them life is discussed in terms meaningful to those involved. The preconceptions of the investigator do not determine the nature of the data obtained. To the degree that social psychology needs to understand the "definition of the situation" of participants, such documents constitute an invaluable source of scientific information.

Because expressive documents are rarely sufficiently controlled by the investigator to afford a crucial test of specific hypotheses, they have generally been used in the exploratory rather than the final stages of the research process. Their greatest value, perhaps, has been in giving investigators a "feel" for the data and thus

producing hunches' with respect to the most fruitful way of conceptualizing the problem. The research scientist must become intimately familiar with the situation under study, and one of the best ways to do this is through careful reading of insightful expressive documents. Thomas and Znaniecki, for instance, conceptualized three types of personality—the Philistine, the Bohemian, and the creative individual—from their documentary materials (52, Vol II, 1853-1859).

Expressive documents are capable not only of identifying the significant variables in a specific problem but of suggesting hypotheses embodying these variables. These formulations will, of course, be influenced by pre-existing theory and thus they are at once inductive and deductive. Sutherland's handling of the Conwell document (16) illustrates this admirably.

The next stage in the scientific process is the verification of hypotheses. Tentative generalizations must be confronted with data sufficient to give them a crucial test. Expressive documents, although they rarely have been, can be used as one of the means of accomplishing this, but they probably can never be solely relied upon because they may not furnish exactly the data necessary in sufficient numbers of cases. Interviews, observation, or questionnaires can provide a better coverage of the pertinent situations because they can be set up to obtain from many cases the information that is precisely relevant to the hypothesis. This is not to say, however, that expressive documents should be completely eliminated at this stage. Just because they are not dominated by the conceptual scheme of the investigator, they constitute an excellent check on data obtained by other methods. They are particularly appropriate to the search for negative evidence.

The use of expressive documents for purposes of nomothetic research is beset with many difficulties and problems. We shall discuss the following six: (1) the optimum scope of a study by this means, (2) the representativeness of the documents, (3) their adequacy, (4) the reliability of the record, (5) the reliability of the analyst's interpretation of the record, and (6) the validity of the findings.

Experience seems to show that a document focused on some specific topic or process is more rewarding nomothetically than one that covers all aspects of a person's life. The smaller the area about

which the informant is writing, the more likely that all the theoretically relevant material will be given in the document. This not only makes possible the framing of specific hypotheses but gives a richness of context that may produce a new conceptual orientation.

Since the investigator is usually trying to arrive at scientific statements that have validity beyond the data from which they are drawn, he must be concerned with the representativeness of the data. This poses a well nigh insuperable problem for one who would work with expressive documents. Even if we allow that the possibility of dictation obviates the exclusion of those who cannot write, there is still the fact that some people are much more interested in expressing themselves than others. It would take the greatest ingenuity to set up a random sample from any universe of persons such that each member of the sample had produced a spontaneous document useful for scientific purposes. Even when the sample is obtained first and the members of it are then stimulated to produce documents to order—which is the best procedure to ensure representativeness—the probability that they will all perform the task satisfactorily is very small. This is perhaps the main reason why expressive documents are thought to be more valuable in the exploratory phases of research than in the definitive testing of hypotheses.

The question of adequacy is not one that concerns only the documents written by the less intelligent. Even highly educated persons may not see all around a problem. They may recount only those aspects of it in which they are interested and leave out altogether aspects of great relevance for social psychological analysis. Moreover, they may not give enough of the background and context to make clear the significance of the behavior described. The provision of a broad topical outline is the best method of coping with this problem, but it can hardly be expected to solve it.

The reliability of the record largely turns upon the truthfulness of the informant. How can one tell whether the author is truthful? A most interesting finding in this connection is that of Frenkel Brunswik: persons rated by associates as unreliable are prone to superlative and absolute statements and to excessive repetition (24). Possibly such a criterion would enable the investigator to reject unreliable documents.

A quite different problem is the reliability of the interpretation made by the social psychologist. How do we know that another scientist would see the same data in the document? This can be, and has been, tested (10, 12, 48). The upshot of these tests is definitely encouraging. Well trained investigators will agree to a scientifically satisfactory degree on what the personality traits and attitudes of persons are and how they are likely to behave under specified circumstances. In order to convince others that this is true, it is highly desirable that as many expressive documents as possible be published so that other investigators can determine to what extent they would agree with the interpretations made.

Lastly, there is the problem of validity. How does one know that the interpretation, even if agreed upon by competent social psychologists, is correct? The only check is to use other criteria, as Cartwright and French (10) and Stouffer (48) have done. The former compared results of the interpretation of case materials including a diary, with actual test performance of the subject and found that the validity of the interpretation of each of two social psychologists was greater than the reliability between their interpretations. What this means is that each man made particularly sound interpretations in certain areas of attitude and behavior but that these same areas were not soundly interpreted by the other. Unreliability among interpreters of documents is not necessarily as damaging scientifically as has usually been thought.

In Stouffer's study four judges rated the attitudes toward Prohibition of 238 writers of biograms. The ratings were validated by correlating them with the subjects' scores on a measured scale. The validity coefficient was 0.81. This is considerably higher than that obtained in the Cartwright and French study but in Stouffer's study the reliability among raters was higher still, 0.96. The higher reliability and validity reported by Stouffer almost certainly resulted from the fact that the biograms and the validating scale were focused on a narrow field—attitudes toward Prohibition. Cartwright and French predicted from expressive documents of a general character how the subject would answer a broad battery of questions comprising a personality test. This would seem to show that valid inferences can be made if the documentary materials for a specific area are adequate. This is perfectly compatible with the

view expressed earlier that a focused document gives the social psychologist the degree of "density of data" necessary to the full employment of his theoretical formulations

This general review of the uses of expressive documents in social psychological research shows that they have genuine value, particularly in the development of concepts and hypotheses, but that for a complete piece of nomothetic research they need to be supplemented by other approaches. Several good studies, such as Thrasher's *The Gang* (53), illustrate combinations of techniques. Over a period of seven years Thrasher directly observed the life of gangs, persuaded youths to write their own stories, interviewed gang members, examined social agency and court records, and clipped newspaper accounts of gang activities. Although he developed hypotheses of a broad type, he did not bring them to a definite test. Pauline Young, in her study of the Molokan sect in Los Angeles (63), likewise employed several approaches. Passive interviews were the source of most of her documents. She also drew on her own observation, newspaper accounts, and public records. At one point she used delinquency statistics to validate a hypothesis that she formulated from the documentary materials.

Although much of the research done with the aid of expressive documents has been rewarding, it is interesting that no study has gone through the scientific process to the point of verifying hypotheses on data different from those with which they were developed. Although this is very expensive and time consuming, we shall not really know the value of this research technique until someone does this.

REGISTRATION AND CENSUS DATA

A great body of statistical data about human populations is collected by government, business, and private agencies. All too frequently, when individual investigators plan their own research, they ignore data available from such sources. Although caution must be observed in the use of these data, they may furnish valuable aids to many phases of social psychological research. A useful classification of such data divides them into registration and census data.

The Nature of Registration and Census Data

Registration data consists of records made at the time of the occurrence of an event in accordance with legal or administrative regulations attached to that event. Such data cover a very wide range of events and comprise a massive record of social life.

The following list is illustrative of important activities covered by registration data.

- | | | |
|----|--|--|
| 1 | <i>Vital events</i> | Births, deaths, marriages, divorces, morbidity |
| 2 | <i>Education</i> | School attendance, grades, performance on psychological tests |
| 3 | <i>Crime</i> | Crimes known to police, arrests, court actions, prison records, parole records |
| 4 | <i>Voting</i> | Registration, voting |
| 5 | <i>Social security payments and benefits</i> | |
| 6 | <i>Automobile registrations</i> | |
| 7 | <i>Draft and Army service</i> | |
| 8 | <i>Illness</i> | Hospital and insurance data |
| 9 | <i>Business activity</i> | Payrolls, production records, absentee records |
| 10 | <i>Formal organizations</i> | Membership, office holding, committee participation |

This list might be extended almost indefinitely, because an essential element in an urban society is an elaborate record keeping system as a basis for knowledge planning and control. A wide variety of events are thus routinely recorded as a normal part of the functioning of the social system. The information about the events may have intrinsic value but its usefulness is greatly enhanced by the collateral information also recorded. For example, records of school enrollment frequently have data on the nativity of parents, occupation of father, place of previous residence for migrants, scores on psychological tests, school grades, etc.

There are a few kinds of registration records which contain direct psychological measures. For example, psychological test results and diagnoses may be a systematic part of the records of schools.

prisons, mental institutions, certain courts, and personnel departments of many business firms. Usually, however, such psychological data are not accumulated in the registration process, but even when they are not, registration data may still be useful in social psychological research.

A census is a periodic collection of data about a population, usually taken by a house to house canvass. In the United States and other urban societies, the census has become the source of great masses of data about human behavior. Although the principal collection agency for this kind of data in this country is the United States Census Bureau, valuable data are also widely collected in school censuses, in real property inventories and in occasional enumerations by other agencies.

The United States Census covers an enormous range both in terms of data collected and the types of geographical units for which they are tabulated.¹ For example, the questions in the 1950 decennial census of population and housing cover such items as age, sex, family composition, labor force status, nativity, amount and source of income, education, migration status, type of dwelling, state of repair of dwelling, rental, etc. For the 1940 census some items of this type were tabulated for units varying in size from the United States as a whole to individual city blocks of large cities. Cross tabulations of many of these variables were published for a variety of geographical units. The census tract data available for small areas of large cities have served as an especially important basis for many social science studies (17, 41, 47). Historical comparisons with earlier decennial census materials may be made for some variables. In a few cases these go back to the first census, made in 1790.

In addition to the decennial censuses of population and housing, significant data are available in the Census of Business and the Census of Manufactures. Data on institutional populations are collected in connection with the population census.

The Census Bureau has recently been collecting and publishing certain current types of data on a sample basis. At the present time these appear under various series of *Current Population Reports*, which include monthly reports on the sizes and composition of the labor force. There are occasional reports on such subjects as family

¹ For a summary of the scope of the population census see Hauser (28). There are occasional guides to available census data (54-58).

composition, migration status, income, education, characteristics of dwelling units, etc. Most of these are national data, but occasionally they are broken down on a regional basis.

In addition to the great mass of published data, it is frequently possible to obtain copies of unpublished tabulated data for the cost of reproduction. Frequently, special tabulations will be made by the Census Bureau at cost, on request. The introductory sections of most publications of the decennial census contain references to the scope of tabulated but unpublished data.

From time to time important methodological and substantive monographs are issued by the Census Bureau. Examples are *State Economic Areas*, by Donald Bogue (59), and *The Growth of Metropolitan Districts in the United States*, by Warren Thompson (55).

Uses of Registration and Census Data

Many of the types of registration data are collected in widely different places and continuously through time. This permits the comparison of the incidence of the recorded events for persons with specific characteristics in different kinds of places and under different temporal conditions. For example, it may be possible to compare birth rates, school attendance, suicide rates, the percentage of the population voting in elections, war bond purchases, etc., as between rural and urban areas, 'poor' and 'rich' areas, and in various phases of the business cycle.

Another use of registration data is in the study of the relationship between nonpsychological variables relevant to a research problem. Social psychological research may then be pursued within the framework provided by the relations between social and situational variables.

A classical example of the use of registration data for relating social variables is a study by Durkheim (19) in which the incidence of suicide was related to membership in certain social groupings. These groupings were then classified by inference according to their social cohesiveness. The result was support for a theory relating the degree of social cohesiveness to the incidence of suicide. Although Durkheim himself did not pursue this line further, a social psychologist given such background material might study this relationship.

with reference to individuals by examining the psychological correlates of social cohesiveness as intervening variables.

A number of sociologists have made studies of deviant behavior in which registration and census data have been used to obtain general sociological relations and personal documents and case records have been used to get at psychological processes relating to such a framework (11, 35). Clifford Shaw (43, 44, 45, 46) followed this pattern in studying both ecological data relating to juvenile delinquency rates and personal-history documents of delinquents.

The possibilities in combining current survey data and registration data have hardly been explored. For example, for a sample of precincts it would be possible to combine survey data on social participation, reference groups, and political activity with registration data on voting, split ballots, registration for voting, campaign expenditures, etc. In this way a study could be made of the relation between attitudes and social organization of local populations and their political behavior.

Another example of the use of registration data is Klineberg's study (31) of selective migration and intelligence. This study depended on school grades, intelligence-test scores, residence histories, and demographic data found in school records.

Closer to the usual interest of the social psychologist is utilization of registration data to select a sample with specified characteristics for a study in which additional data will be gathered. From the point of view of experimental design, this may have two purposes: (1) to select a number of groups which are similar with respect to important characteristics; such "equated" groups may then be divided into control and experimental groups to eliminate the equated characteristics as variables in the study; (2) to select groups which differ on a characteristic which is a variable under study in the research.

The first purpose is illustrated by many studies (27, pp. 57ff.) with school children in which they are divided into control and experimental groups on the basis of school records. Similarly, in the film experiments reported in *Studies in Social Psychology in World War II* (29), the control and experimental groups were chosen by matching military units which had roughly similar characteristics according to Army records.

The second purpose is illustrated by investigations in the field of industrial relations in which plant records of the worker's productivity, demographic characteristics, and work history may be related to current survey or observational data (15, 22)

Another type of research is illustrated in the study by Chapin and Jahn (14, pp 41-50) in which data from WPA and relief records were used in conjunction with morale measurements made by the investigators to study the effect of type of relief on the morale of the recipient. Studies of the effect of certain types of appeals on war bond sales (9) also depended on the records of war bond sales as the criterion variable. In these various types of studies at least one of the principal variables was measured with registration data while another was measured directly by the investigator.

Whenever, as is usually the case, a field research investigation has a geographical context, census publications provide a valuable source of descriptive material about the population in the area involved.

Chapter 5 has already indicated that census data are an important basis for the construction of area samples. Census data may also be used to select areas or communities as the units for research study. In the study *Public Response to Peacetime Uses of Atomic Energy* (30) a group of communities near atomic energy installations were matched on census characteristics with another sample of communities removed from such installations. Then certain attitudes of the members of the paired communities were compared.

Like registration data, census data may be used to study the relationships among social and economic variables relevant to a problem. Such data generally do not include direct measures of psychological variables. However, the relationships among background variables are frequently important for the social psychologist, and they may facilitate interpretation of other data. For example, in a study of social psychological factors affecting home purchases (49) it was useful to refer, in analysis, to current census data on the tenure status of recent migrants.

Unlike registration data, it is not generally possible to link census data for individuals directly to other social and psychological data about these individuals, because the Census Bureau must keep information about individual respondents completely confidential.

However, through "matching" studies it is sometimes possible to combine census and other data as follows (36, pp 14 15)

The Bureau of Census will compile statistics from specific census schedules of lists of consumers furnished by business concerns. A given firm may provide certain information about its clients which is transcribed and placed on punch cards along with information taken from the census schedules of the same persons. Cross tabulations are then prepared to show relationships between the characteristics known to the census and those known to the enterprise. No data are given out in terms of individuals but only in frequencies or summary form.

Historical Series of Registration and Census Data

For general background purposes as well as for analyses relating different time series of events, historical series of registration and census data may be important. The use of time series in economic research is highly developed. A considerable number of sociological studies have related time series of social events to economic cycles, wars, and other crisis events.

Although time series involving direct measurements of psychological variables are rare, there have been comparisons over time based on asking the same attitudinal questions in several surveys (8, pp 220 232). Such series are likely to become more important as the number of annual surveys with some continuity in questions increases. For example, such data are beginning to accumulate for psychological variables affecting consumer decisions in the annual Survey of Consumer Finances conducted by the Survey Research Center of the University of Michigan for the Federal Reserve Board.² Similarly, a number of survey organizations have been collecting comparable attitudinal data about foreign policy in successive surveys. It is likely that such time series will grow in number. Their importance will be greatly enhanced when they cover a sufficient time period to make it possible to relate them to established time series of census and registration data.

The extent of the historical statistical series already available

² Reports on these surveys appear frequently in *The Federal Reserve Bulletin*.

The second purpose is illustrated by investigations in the field of industrial relations in which plant records of the worker's productivity, demographic characteristics, and work history may be related to current survey or observational data (15, 22)

Another type of research is illustrated in the study by Chapin and Jahn (14, pp 41 50) in which data from WPA and relief records were used in conjunction with morale measurements made by the investigators to study the effect of type of relief on the morale of the recipient. Studies of the effect of certain types of appeals on war bond sales (9) also depended on the records of war bond sales as the criterion variable. In these various types of studies at least one of the principal variables was measured with registration data while another was measured directly by the investigator.

Whenever, as is usually the case, a field research investigation has a geographical context, census publications provide a valuable source of descriptive material about the population in the area involved.

Chapter 5 has already indicated that census data are an important basis for the construction of area samples. Census data may also be used to select areas or communities as the units for research study. In the study *Public Response to Peacetime Uses of Atomic Energy* (50), a group of communities near atomic energy installations were matched on census characteristics with another sample of communities removed from such installations. Then certain attitudes of the members of the paired communities were compared.

Like registration data, census data may be used to study the relationships among social and economic variables relevant to a problem. Such data generally do not include direct measures of psychological variables. However, the relationships among background variables are frequently important for the social psychologist and they may facilitate interpretation of other data. For example, in a study of social psychological factors affecting home purchases (49) it was useful to refer, in analysis, to current census data on the tenure status of recent migrants.

Unlike registration data, it is not generally possible to link census data for individuals directly to other social and psychological data about these individuals, because the Census Bureau must keep information about individual respondents completely confidential.

However, through "matching" studies it is sometimes possible to combine census and other data as follows (36, pp 14 15)

The Bureau of Census will compile statistics from specific census schedules of lists of consumers furnished by business concerns. A given firm may provide certain information about its clients which is transcribed and placed on punch cards along with information taken from the census schedules of the same persons. Cross tabulations are then prepared to show relationships between the characteristics known to the census and those known to the enterprise. No data are given out in terms of individuals but only in frequencies or summary form.

Historical Series of Registration and Census Data

For general background purposes as well as for analyses relating different time series of events, historical series of registration and census data may be important. The use of time series in economic research is highly developed. A considerable number of sociological studies have related time series of social events to economic cycles, wars, and other crisis events.

Although time series involving direct measurements of psychological variables are rare, there have been comparisons over time based on asking the same attitudinal questions in several surveys (8, pp 220 232). Such series are likely to become more important as the number of annual surveys with some continuity in questions increases. For example, such data are beginning to accumulate for psychological variables affecting consumer decisions in the annual Survey of Consumer Finances conducted by the Survey Research Center of the University of Michigan for the Federal Reserve Board.² Similarly, a number of survey organizations have been collecting comparable attitudinal data about foreign policy in successive surveys. It is likely that such time series will grow in number. Their importance will be greatly enhanced when they cover a sufficient time period to make it possible to relate them to established time series of census and registration data.

The extent of the historical statistical series already available

² Reports on these surveys appear frequently in *The Federal Reserve Bulletin*.

is indicated by the fact that approximately 3000 statistical time series covering various periods from 1789 to 1945, have been published in the volume *Historical Statistics of the United States 1789-1945* (57). An appendix in the *Statistical Abstract of the United States* now brings many of these series up to date.

The following statement from the census volume on historical statistics (57 p. vi) serves to emphasize the cautions already expressed in connection with the use of such historical series

identification of changes in concept and coverage over a period of time is important since such changes may affect vitally the interpretation of the statistics for a span of years. Coupled with this is the need for definitions of terms employed in published historical tables definitions which may be in a separate publication or may never have been published.

The Use of Published Indices

There are many useful published indices based on the combination or other manipulations of registration and/or census data. Illustrations are the Cost of Living Index, various indices of business conditions, county level-of-living indices, juvenile-delinquency rates, the F B I crime rate index, offenses known to the police, vital rates (birth, death, etc.).

The conflict between the goals of historical comparability and current validity is a general problem of such historical series and is likely to be especially important in the case of constructed indices. This problem may be illustrated with reference to the Cost of Living Index (60) which is essentially a measure of changes in the cost of a fixed list and quantity of consumer goods. The weighting of the items was originally based on a survey of the consumption of wage earners and clerical workers in 1917-1919. The items and weights were revised on the basis of a study of expenditures of moderate income families in 1950. For purposes of comparison over time it would be desirable to keep the items and their weights fixed so that only the cost of buying this fixed list of goods would vary. However, from the point of view of current validity, this might result in an unrealistic index in which items and their weights did not represent current consumption habits. In practice a compromise

is made between the conflicting goals of comparability and current validity. The items and weights are revised from time to time but not each year. When the revision is made, both the old and the new series are computed for a while for linkage purposes.

It is frequently possible for the individual investigator to construct indices for a specific purpose by combining or manipulating series of published data. Although such indices usually do not meet rigid scaling standards, they are frequently useful as rough measures of a variable. Examples of such indices based on registration and census data are those for "plane of living" (25), "moral integration" (6), and "segregation" (30). Ingenuity on the part of the researcher frequently enables him to use such indices as rough measures of social variables. In many field situations the range of variation is so great that even rough indices will serve to differentiate at a satisfactory level.

The Ex Post Facto Design³

Registration and census data frequently provide data for contrasting groups which have already been differentially subject to some stimulus. Studies based on such data have become known as *ex post facto* studies. Greenwood (27) defines the *ex post facto* experiment as one in which "we work backward by controlling after the stimulus has already operated, thereby reconstructing what might have been an experimental situation." This is to say that the stimulus is not controlled by the investigator. The nature of the experimental manipulations the investigator can make are strictly limited.

The Chapin and John study (14) of the relation of type of relief to worker's morale, cited earlier in this chapter, is an example of an *ex post facto* study based on registration data. In this case the stimulus (type of relief) was not controlled by the investigator. A group of persons receiving work relief through the WPA was matched with a group receiving direct relief but eligible for WPA work relief. The characteristics used for matching were age, sex, race, nativity, amount of education, usual occupation, size of family, and length of time on relief. Morale measurements were made on both

³ Students interested in an intensive study of the *ex post facto* design should consult Greenwood's (27) careful analyses.

groups. The group receiving work relief scored significantly higher on morale tests than the group on direct relief.

An example of ex post facto research based on census schedule data is the study by Freedman and Hawley (23) of the relation between unemployment and migration. In this case the migrants in two Michigan cities were matched on a number of relevant census characteristics with nonmigrants at their place of origin. The two groups were then compared on unemployment rates prior to migration. The premigration unemployment rate for the migrants was slightly higher than that for nonmigrants. The difference was much less than had been expected on the theory that unemployment was an important cause of migration for comparable groups.

A large number of other suggestive ex post facto studies have been done without close matching on control variables. The largest number of empirical sociological studies are of this character. An excellent study which made ingenious use of registration data to investigate an important problem is that by A. J. Reiss, Jr. (38) of the relation between juvenile delinquency and the failure of personal and social controls. This study is based on the case materials including psychiatric diagnoses and the case records of social workers, for 1110 juvenile delinquent probationers in the official juvenile court records. On the basis of his analysis of these data Reiss reached the following conclusions, which are illustrative of types of important relationships which may be investigated by various ex post facto designs:

Our observations show (1) that delinquent recidivists are less often than non recidivists members of social groups and live in a social milieu which is characterized by norms and effective techniques in producing conformity behavior contra delinquency (2) that delinquent recidivists less often *accept* or submit to the control of social groups which enforce such conformity behavior than do non recidivists and (3) that delinquent recidivists are less often persons with mature ego ideals or non delinquent social roles and appropriate and flexible rational controls which permit the individuals to guide action in accord with non delinquent social group expectations (38 p. 204)

The most important limitation of ex post facto studies of this kind is that the members of the control and experimental groups

are 'self selected' rather than randomly assigned. This means that the differences found between the control and experimental group may be due to factors connected with characteristics of these groups other than the experimental stimulus or the characteristics matched. For example, in the Chapin and Jahn study (14) such factors as persistence or political activity may conceivably explain both the stimulus variable (type of relief) and the effect variable (morale). Despite the matching process, it is possible that the two groups differed in morale even before they were subjected to different relief programs. Similar uncontrolled selective variables may account for the results of the Freedman Hawley study (23). Even with this limitation, the ex post facto design may be useful in field studies, especially if replication under various conditions is possible. It is important, however, to understand the limited level of generalization which is usually possible with this type of design.

Problems in the Use of Registration Data

The fundamental limitations of registration data for social psychological research arise from the fact that they are not generally collected for the specific purposes of such research. The definitions and tabulations used in calculating and processing the data may differ from those which the researcher would use in collecting data for his own purposes. For example, the detail with which occupational data are recorded in school files may not be ideal for the investigator who wishes to use occupation as an index of social class. Most of the limitations of registration data come from the basic fact that the investigator cannot impose his own standards of validity and reliability on the data.

The completeness of coverage of registration data varies from time to time in accordance with the efficiency of the data collection, the nature of the data and the incentives which the population has to record the event involved. For example, it has frequently been observed that a recorded rising birth rate in a country undergoing modern development is not necessarily proof that the birth rate is actually rising. It is more likely to indicate that the efficiency of birth registration has improved along with other statistical services.

Similarly, variations in the reported incidence of some registered event may reflect variations in the vigor of data collection.

rather than variations in the event itself Kuczynski (32 p 8) reports that extreme variations in vital rates of certain countries resulted from periodic campaigns to register events not recorded in previous periods The extreme irregularities in these rates were largely statistical artifacts

Because registration data are collected as an incident of administrative processes they may suffer from their particular context Income tax returns have obvious limitations as accurate measures of income in many countries because of the motivations for tax evasion Robison (39) has pointed out that the chance that a delinquent act will be officially recorded is related to the social status and ethnic background of the delinquent Similarly Sutherland's study (51) of white collar crime indicates that certain crimes committed by persons of high social status less frequently reach certain recorded stages of the police or judicial system

In evaluating the validity of data the investigator should consider whether the respondent is able to give the information required and whether the recorder has a strong motivation to secure the information accurately Lundberg (33 p 135) reports that at various times the cause of death on death certificates has been recorded as died suddenly nothing serious death caused by five doctors vital statistics These were cases in which the report was made by unqualified lay persons Even qualified personnel are limited by the state of knowledge in the field Increases in death rates from certain causes are believed to be partly attributable to an increased skill in diagnoses of these causes by the medical profession

In some cases the informant has the necessary information but the recorder is not motivated or properly instructed to elicit it For example where ethnic background is obtained as incidental to the record of some event the recorder may feel that the data have no relevance for his agency or he may be improperly trained to obtain and record the data Fortunately in an increasing number of cases agencies are training recorders of data with instructions parallel to those given to the interviewer for the use of a schedule

In using time series of registration data it is important to be sure that regulations covering the records have not changed Obviously comparisons of the number of income tax payers over a period of time is no indication of income levels if the minimum income subject to taxation has changed

A special problem of this kind may be changes in the practice of assigning the event to the place where it occurs or to the legal residence of the person to whom it occurs. The registration of births and deaths is a case in point. At the present time these data are tabulated both as to place of occurrence and place of legal residence. That such allocation of the events is an important issue may be seen from the fact that in 1940 21 percent of all registered births were coded "nonresident." This percentage varied from 8 percent for rural places to 40 percent for births in cities of 2500 to 10,000 in population (61, p. 17).

In some cases the reporting agency may have made studies to test the reliability or validity of its data. In other cases, the investigator himself should make such tests whenever possible by checking for internal consistency in the data or by comparison with other series. One indication of the coverage of a registration series is the 1950 check comparison of birth registration with infants, reported in the 1950 U. S. Census (42). For the United States as a whole, birth reporting was estimated to be 97.8 percent complete. Estimates for individual states ranged from 88.1 percent for Arkansas to 100.0 percent for Connecticut. It was less than 95 percent in only seven states. The 1950 figure of 97.8 percent for the United States as a whole represents marked improvement from the 1940 figure of 92.5 percent.

The extent to which registration data are subject to the limitations we have presented varies widely with the administrative context in which they are collected. In some cases, the collection is intelligently and specifically guided by the research purposes of the agency or for general research purposes. However, it is obviously important that the researcher who uses registration data should know something of the special definitions and contexts which may affect the nature of the data collected. The fact that data are published under official auspices does not guarantee their quality.

Problems in the Use of Census Data

To a certain extent, the census, as a source of research data, is subject to the same limitations as registration. Most important, the definitions used in data collection may not always be completely appropriate for the purposes of a specific piece of research. It is

also true that over a period of time changes are made in the classification system and these may produce errors in interpretation if they are not clearly understood. For example, between the 1930 and 1940 census the basis for identifying the working population was changed. In 1930 and earlier years, the working population was identified with gainful workers—persons having a gainful occupation regardless of whether they were actually working or seeking work at the time of the census, since 1940 the 'labor force' concept has been used to classify workers on the basis of their activity in a specified period at or immediately preceding the census.

The importance of understanding the working definitions by which data are collected is illustrated in a recent sample survey by the Census Bureau (56). In this survey 844,000 persons classified as having moved between farm and nonfarm residences had not actually changed their physical location. Their classification as migrants resulted from a change in the use of the land on which they were resident. Similarly, estimates of the amount of unemployment based on sample census data vary according to the definition of what constitutes part time employment, the status of unpaid family workers and the line of demarcation between those who are unemployed and those not in the labor force. These illustrations document the common sense point that terms found in published tables should not be taken at face value. A careful study should be made of the definitions usually found in census volumes and the instructions to enumerators for collecting data.

In general, the data of recent U. S. censuses are of a relatively high level of reliability and validity. The personnel of the Census Bureau includes an outstanding group of statisticians and social scientists. Within the limitations of an operation of its magnitude, the Census Bureau provides a model in many respects for methodological aspects of survey work. Where known sources of error exist, they are usually pointed out in census publications. There has been an increasing effort to incorporate checks on the quality of the data into the operations of the Census Bureau (34).

SUMMARY

This chapter has dealt with the use in research of certain types of data collected by persons other than the investigator himself. Such data are available in great quantity and for a wide range of

important problems as a result of the elaborate record keeping and documentation in our society. The principal limitation of such data is that the operational definitions of the data and the possibilities of experimental manipulation are outside the experimenter's control. Although this restricts their usefulness, such data remain very important. They provide unique access to historical social situations and to some current social situations which are otherwise difficult or expensive to observe. Moreover, these are data in a "natural" social setting.

BIBLIOGRAPHY

- 1 Abel T *Why Hitler came into power* New York: Prentice Hall, 1938.
- 2 ——— The nature and use of biograms *Amer J Sociol*, 1947, 53, 111-118.
- 3 Allport G W *The use of personal documents in psychological science* New York: Social Science Research Council, 1942.
- 4 ———, Bruner, J S, and Jandorf, E M Personality under social catastrophe *Character and Pers*, 1941, 10, 1-22.
- 5 Angell R C *The family encounters the depression* New York: Scribner, 1936.
- 6 ——— The moral integration of cities *Amer J Sociol* 1951, 47, No. 1, Part 2, 1-140.
- 7 Blumer, H *An appraisal of Thomas and Znaniecki's 'The Polish peasant in Europe and America'* New York: Social Science Research Council, 1939.
- 8 Cantril H *Gauging public opinion* Princeton: Princeton Univ. Press, 1944.
- 9 Cartwright D Some principles of mass persuasion *Hum Relat* 1949, 2, 253-268.
- 10 ——— and French J R P Jr The reliability of life history studies *Character and Pers* 1939, 8, 110-119.
- 11 Cavan R S *Suicide* Chicago: Univ of Chicago Press, 1928.
- 12 ———, Hauser, P M, and Stouffer, S A A note on the statistical treatment of life history material *Soc Forces*, 1930, 9, 200-203.

- 13 Chłelasinski J *The young peasant generation in Poland* (in Polish) Warsaw State Institute of Rural Culture 1938
- 14 Chapin F S *Experimental designs in sociological research* New York Harper 1947
- 15 Coch L and French J R P Jr *Overcoming resistance to change* *Hum Relat* 1948 1 512-532
- 16 Conwell C and Sutherland E H *The professional thief, by a professional thief* Chicago Univ of Chicago Press 1937
- 17 Cressey P F *The succession of cultural groups in the city of Chicago* Unpublished Ph D thesis Univ of Chicago 1930
- 18 Dollard J *Criteria for the life history* New Haven Yale Univ Press 1935
- 19 Durkheim E *Suicide a study in sociology* Glencoe Free Press 1951
- 20 Dyk W *Son of Old Man Hat* New York Harcourt 1938
- 21 Ford C S *Smoke from their fires* New Haven Yale Univ Press 1941
- 22 Fox J B and Scott J F *Absenteeism* *Publ Grad Sch Bus Adm Harv Univ* 1943 No 29
- 23 Freedman R and Hawley A *Unemployment and migration in the depression* *J Amer Stat Assoc* 1949 44 260-272
- 24 Frenkel Brunswick E *Mechanisms of self deception* *J Soc Psychol* 1939 10 409-420
- 25 Goodrich C *Migration and economic opportunity* Philadelphia Univ of Pennsylvania Press 1936
- 26 Gottschalk L Kluckhohn C and Angell R *The use of personal documents in history, anthropology and sociology* New York Social Science Research Council 1945
- 27 Greenwood E *Experimental sociology, a study in method* New York Kings Crown 1945
- 28 Hauser P M *Population* In Hauser P M and Leonard W R (eds) *Government statistics for business use* New York Wiley 1946 pp 325-358
- 29 Hovland C I Lumsdaine A A and Sheffield F D *Experiments on mass communication* Princeton Princeton Univ Press 1949
- 30 Jahn J Schmid C F and Schrag C *The measurement of ecological segregation* *Amer Sociol Rev*, 1947 12, 293-303
- 31 Klueberg O *Negro intelligence and selective migration* New York Columbia Univ Press 1935
- 32 Kuczynski R *The measurement of population growth* New York Oxford 1936

- 33 Lundberg G *Social research* New York Longmans Green 1929
- 34 Marks E S and Mauldin W P Response errors in census research
J Amer Stat Assoc, 1950 45 424 438
- 35 Mowrer E R *Family disorganization* Chicago Univ of Chicago Press 1939
- 36 Parten M B *Surveys polls and samples* New York Harper 1950
- 37 Radin P *Crashing Thunder* New York Appleton Century 1926
- 38 Reiss A J Jr Delinquency is the failure of personal and social controls
Amer Sociol Rev 1951 16 196 207
- 39 Robison S *Can delinquency be measured?* New York Columbia Univ Press 1936
- 40 Runner J R Social distance in adolescent relationships
Amer J Sociol 1937 43 428 439
- 41 Schmid C F *Social trends in Seattle* Seattle Univ of Washington Press 1944
- 42 Shapiro S and Schachter J Birth Registration Completeness United States 1950
Public Health Reports, 1952 6, 513 524
- 43 Shaw C R *The jack roller a delinquent boys own story* Chicago Univ of Chicago Press 1930
- 44 ——— *The natural history of a delinquent career* Chicago Univ of Chicago Press 1931
- 45 ——— *Brothers in crime* Chicago Univ of Chicago Press 1938
- 46 ——— and McKay H *Juvenile delinquency and urban areas a study of rates of delinquents in relation to differential characteristics of local communities in American cities* Chicago Univ of Chicago Press 1942
- 47 Shevsky E and Williams M *The social areas of Los Angeles* Berkeley Univ of California Press 1949
- 48 Stouffer S A *An experimental comparison of statistical and case history methods of attitude research* Unpublished Ph D thesis Univ of Chicago 1930
- 49 Survey Research Center (Institute for Social Research) *Relevant considerations in recent home purchases* Ann Arbor Survey Research Center 1950
- 50 ——— *Public response to peacetime uses of atomic energy* Ann Arbor Survey Research Center 1951
- 51 Sutherland E H *White collar crime* New York Dryden Press 1919
- 52 Thomas W I and Znaniecki F *The Polish peasant in Europe and America* New York Knopf 1927
- 53 Thrasher F M *The gang* Chicago Univ of Chicago Press 1927

- 51 U S Bureau of the Census *Topical index of population census reports 1900 1930* Ann Arbor Edwards 1931
- 52 ——— *The growth of metropolitan districts in the United States 1900 1940* (by Warren Thompson) Washington U S Govt Printing Off 1947
- 56 ——— *Internal migration in the United States April 1940 to April 1947 Current Population Reports 1948 Series P 20 No 14 1 29*
- 57 ——— *Historical statistics of the United States 1789 1945* Washington U S Govt Printing Off 1949
- 58 ——— *Catalog of United States census publications 1790 1945* Washington U S Govt Printing Off 1950
- 59 ——— *State economic areas a description of procedure used in making functional grouping of counties of United States* (by Donald Boguc) Washington U S Govt Printing Off 1951
- 60 U S Bureau of Labor Statistics *The revised Consumer Price Index Monthly Labor Rev 1953* 76, 161 175
- 61 U S National Office of Vital Statistics *Vital statistics rates in the United States 1900 1940* (by Forrest E Linder and Robert D Grove) Washington U S Govt Printing Off 1947
- 62 Wells H G *Experiment in autobiography* New York Macmillan 1931
- 63 Young P V *The pilgrims of Russian town* Chicago Univ of Chicago Press 1932

The Collection of Data by Interviewing

Charles F. Cannell and Robert L. Kahn

In almost every field of human thought it is possible to observe indications of the laborious ascent from superstition and mysticism to scientific fact. Such observations reveal that improvement in the systematic collection of data is a major characteristic of scientific development. In the long established physical sciences the instruments and techniques of data collection are well developed, in the social sciences the development of techniques for measurement and quantification has recently become a focus of effort and attention.

To some extent the needs of the social sciences for data can be met through techniques of observation and physical measurement. To an increasing degree, however, social science is demanding data which must be reported by individuals out of their own experience. Attitudes, perceptions, expectations, anticipated behavior, are available to the economist, sociologist, psychologist, and anthropologist only through such direct communication.

In a sense, of course, social scientists have always "communicated" with people and derived insights from such communications. The problem for social science is to transform the highly subjective process of "getting insights" into a systematic method for the collection of social data. This chapter discusses some of the principles and

techniques by which the process of interviewing can be made to approach the criteria for scientific measurement

Criteria for Scientific Data Collection

The adequacy of a technique for collecting data is ordinarily judged in terms of criteria of reliability and validity, concepts which are discussed at length in Chapter 6. Reliability requires that repeated measurements yield results which are identical or fall within narrow and predictable limits of variability. The criterion of validity demands that the measurement be meaningfully related to the research objectives—that is, that it measure what it purports to measure.

Both these criteria apply not only to the data collection instrument but also to the technique and procedure specified for using the instrument. The reliability and validity of social data depend not only on the design of the questionnaire or interview schedule but also upon the manner of administering the instrument, the technique of interviewing. The techniques discussed in this chapter for wording questions, constructing questionnaires, and conducting interviews are attempts to aid the researcher in approximating the twin goals of reliability and validity in his data collection.

Potentialities of the Interview

We are concerned here with the interview as a device for collecting data required to test hypotheses in social research. The principles which govern questionnaire design, interviewing, and the training of interviewers are, however, relevant to most situations in which information is desired from a respondent. Thus, the lawyer must interview his client in order to represent or defend him, the physician must base his diagnosis upon the medical interview as well as the examination, the journalist, the personnel officer, the social worker—all depend to some extent upon their skills as interviewers as well as upon those other skills which their professions demand of them.

The fact that the interview is used very widely does not imply that it is the best device for collecting social data in all circumstances. One of the choices which the social scientist must make

involves the various methods of acquiring data. One of the important criteria is the relative accessibility of the required data to different means of collection. The sources and disposition of family income provide an example of data which are at present virtually inaccessible from sources other than personal interview.

Suppose, for example, that our research objective is to test some hypotheses about the relationship between the source and amount of family income and the pattern of saving and spending. This objective requires that we assemble data on income and expenditures for individual families. Although gross expenditures for different commodities might be estimated from data provided by manufacturers or trade establishments, and the volume of savings might be determined from banks, the pattern of income and expenditure of family units cannot be reconstructed from external sources. Such information is uniquely available through interviews with a sample of family units.

On the other hand, there are many data relating to income and expenditure which can be obtained with accuracy and economy by means other than the interview survey. We might, for example, wish to test the hypothesis that sales of government bonds through pay roll deduction plans tend to increase after a plant wide wage adjustment. If this is our research objective, it is likely that the company records, perhaps in combination with those of the Treasury Department, can best meet our need for data. The alternative of interviewing relatively large numbers of people in order to identify the few bond buyers and then questioning this group about its rate of purchase and recent fluctuations in income is costly and complicated by comparison.

Another kind of data which has been successfully collected by means of interview and personally administered questionnaires has to do with the attitudes, perceptions, and behavior of people in work situations. For example, we might study the hypothesis that a worker's motivation to produce will be related to the intrinsic satisfaction which he derives from his job. Or we may hypothesize that worker productivity depends upon the individual's perceptions of the consequences of high or low productivity, and the extent to which these consequences represent personal goals for him. Either of these hypotheses requires data which are 'inside the individual' and which he alone is capable of communicating. Any other ap-

proach to assessing the individual's satisfaction with his job would almost certainly involve a risky process of deduction and inference.

Even when the research objectives call for information which is beyond the individual's power to provide directly, the interview is often an effective means of obtaining the desired data. The studies of prejudice and ethnocentrism by Adorno *et al* provide an example of such research.¹ The research design called for the rating of individuals along a number of dimensions, including anti-Semitism and other ethnocentric characteristics, politico-economic conservatism, and several aspects of personality organization.¹ Bias and lack of training make it impossible for an individual to provide directly and with validity such intimate information about himself, even if he is motivated to the utmost frankness. But only he can provide the data about his attitudes toward his parents, colleagues, and members of minority groups, from which some of his deeper lying characteristics can be inferred.

In short, if the focal data for a research project are the attitudes and perceptions of individuals, the most direct and often the most fruitful approach is to ask the individuals themselves. As Jahoda, Deutsch, and Cook (11) suggest, observational methods are of primary value in describing and studying behavior which takes place in a controlled situation, in response to known stimuli. Observational methods are less likely to be useful for the measurement of attitudes and perceptions and are obviously unable to probe the past or to determine an individual's intentions for the future. The criteria of directness and economy, and the ability to collect data about beliefs, feelings, past experiences, and future intentions have widened the range of application of the interview. The interview, however, is not without its own limitations.

Limitations of the Interview

One of the limitations of the interview is the involvement of the individual in the data he is reporting and the consequent likelihood of bias. Even if we assume the individual to be in possession of certain facts, he may withhold or distort them because to

¹ In this series of studies interviews were used for exploratory purposes for the development of hypotheses and instruments and for validation of data obtained by written questionnaires.

communicate them is threatening or in some manner destructive to his ego. Thus, extremely deviant opinions and behavior, as well as highly personal data, have long been suspect when obtained by personal interviews. Nevertheless, much experience in recent years indicates that such limitations on interview subject matter are not to be rigidly assumed.

Another limitation on the scope of the interview is the inability of the respondent to provide certain types of information. For example, the hypothesis that paranoid tendencies are related to an inability to work with groups demands some measurement of the respondent's personality structure. Although he may be completely unqualified to make a direct judgment of such characteristics in himself, he is uniquely qualified to provide some personal information from which an expert might make a diagnosis. Thus, the inability of the respondent to provide a certain datum may mean that a different means of data collection is advisable or it may mean that the interview must be so constructed that the respondent provides raw data which are relatively available and nonthreatening to him so that experts may then interpret his responses in order to provide the information specified by the research objectives.

Memory bias is another factor which renders the respondent unable to provide accurate information. Often, the only clear way around the problem of recall is to have the foresight and facilities to carry out a research design over a period of time applying appropriate measurements at the time intervals indicated by the research objectives.

Summary

In summary, the interview and questionnaire appear as powerful instruments for social research, and the range of their usefulness is steadily widening. Individuals' past experiences and future behavior are virtually unobtainable by other means. Perceptions, attitudes, and opinions which cannot be inferred by observation are accessible through interviews. The major problems in interviewing stem from the inability or unwillingness of the respondent to communicate. These problems, as we have seen, can be surmounted wholly or in part by various means. The skills and technique of the interviewer, the ingenuity of the data collecting instrument and

the knowledge of the analyst can compensate to some degree for the biases, memory failures and inexpertness of the respondent

THE PSYCHOLOGICAL BASIS OF THE INTERVIEW

In discussing the process of collecting social data, we have implied that the questionnaire is a measuring instrument or device used by the social scientist in much the same way that specialized instruments of measurement are utilized in other fields. The interviewer is a technician who manipulates the instrument, takes the appropriate readings, and records the results. In this sense the interviewer's function parallels that of scientific technicians in other fields. First, he must be provided with a questionnaire which is adequate to the research objectives. Secondly, he must ask the questions and record the responses in a standard way.

Considering the interviewer as a scientific technician and the interviewing process as a scientific technique implies that we are able, through the application of a specific instrument in a specific manner, to achieve identical results in given situations. There is further implication that we are able to specify explicitly each step which the technician must follow in using the instrument.

There is sufficient similarity between the scientific technician and the interviewer to make the analogy attractive; however, it does not hold completely. If the interviewer were only to ask a specific question in a standard way, he would not succeed in obtaining responses from different respondents which reflected the same degree of frankness, the same amount of completeness, and so on. In short, the interviewer cannot apply unvaryingly a specified set of techniques, because he is dealing with a varying situation.

The chemical technician, whatever the complexities of the substance with which he must deal, is not confronted with the necessity for taking into account the defenses, the varying motivations, and the diverse perceptions of the substance with which he is dealing. The interviewer, on the other hand, must take account of such social psychological manifestations as these, and the measure of his success as an interviewer is very largely dependent upon the extent to which he is insightful and successful in recognizing and

dealing with the social-psychological phenomena of the interviewing process.

Contemporary social science does not provide the interviewer with adequate methods for dealing with all the variables at work in the interview. To some extent this might be thought of as a symptom of the youthful inadequacy of social science in general and social psychology in particular. To a considerable extent, however, it is also a function of the unusual complexity of the subject matter on which the interviewer, as a scientific technician, is exercising his techniques.

Much of the available literature consists of rules of thumb, presented as lists of "do's" and "don't's" for the interviewer and for the questionnaire framer. These do's and don't's are essentially non-systematic compilations of interviewing experience derived from a variety of situations over a considerable period of time. One might regard them as the "folklore" of interviewing, based on experience, and thus as having a good deal of pragmatic utility. They often represent practices which have achieved a degree of success in a variety of situations. Nevertheless, they have the disadvantage of being somewhat unsystematic in their approach to the interview process. There are few scientific studies testing these common-sense injunctions, and one must accept or reject, without proof, most of those interviewing principles which other people have judged to work. A final disadvantage of the common-sense rules for interviewing is that at best they represent a superficial statement of what constitutes successful interviewing procedures. They do not help us to understand the interpersonal relations between interviewer and respondent. They do not tell us why a specific practice makes for a successful or unsuccessful interview, or within what range a specific practice is desirable or undesirable.

Until we have a theoretical basis for understanding the interview process, and until we have tested empirically some of the interviewing folklore which we frequently take for granted, we are unlikely to advance in a basic way our knowledge and practice of interviewing procedures. In other words, we have every reason to suspect that we possess a powerful instrument for collecting research data, but we do not yet know its full potentialities and limitations.

Unfortunately, social science has not yet provided a comprehen-

tion, although insufficient for an adequate interview, at least enables the interviewer to describe the project and permits him to take the first steps toward building sufficient motivation to permit the interview to move forward. The initial reactions of respondents vary considerably among different segments of the population and these variations are reflected in the refusal rates recorded by various survey agencies. For example, the higher rate of initial refusals in urban areas illustrates a difference between urban and rural populations with respect to attitudes toward the casual caller and the appropriate behavior toward him.

For some parts of the population, the customary manner of reacting to authority figures may do much to determine the initial reaction to an interviewer. Thus, the interviewer may gain access to a respondent because he is perceived as an individual or as a representative of an agency possessing authority and commanding respect from the respondent. For example, a respondent may react favorably to the interviewer because he represents a well known research organization, a university, or a governmental agency.

The interviewer necessarily accepts these motives as a basis for beginning to communicate with the respondent. However, he immediately begins to define the situation in a manner which relates the interview to certain goals the respondent is suspected to cherish and accordingly, gives the interview a positive valence for the respondent.

In some studies, the relating of the interview to the respondent's goals may be started in advance of the interview itself by means of letters, radio, or newspaper announcements. The interviewer's introduction and statement of the purpose of the research also are intended to make the interview appear compatible with or even as a means of achieving, some respondent goal.

After initial acceptance by the respondent, the interview begins with questions designed to develop active interest on the part of the respondent. These are the kinds of items often referred to as 'rapport builders'. The purpose of such questions is to motivate the respondent by assuring him that the interview will be interesting—that is, that its content is related to interests or goals which he already has. An additional purpose of these introductory questions is to relieve anxieties which the respondent may entertain in regard to his own ability to play the respondent role effectively. This is

done by educating him in, or clarifying for him, the type of response expected, thus demonstrating to him his ability to handle the prescribed role. If the initial phase of the interview is successful, the interview has reached the point where one of two major types of motivation may be tapped, thereby ensuring continued cooperation by the respondent.

Recent work in communications within small groups has resulted in findings which appear relevant to our understanding of communication between an interviewer and a respondent. These findings can be summarized as follows. One of the motives for communicating is the desire to influence, in some manner, the person to whom the communication is addressed. That is, a person will communicate in a given situation if he believes that such communication will bring about a change or effect an action which he considers desirable (9). In the interview situation, this means either that the interviewer will be perceived as a person who can bring about change directly or that the interview will be seen as a potential vehicle for indirectly accomplishing a desired change. A clinical psychologist, a social worker, or physician is frequently perceived by the patient or client as a direct agent of change. The respondent is likely to feel that communicating his symptoms or his financial difficulties to the professional person will benefit him directly. An example of a less direct relationship between interviewer and the goal of a respondent is provided by the typical market-research survey, in which the respondent believes that by his expressed preference for a specific kind of packaging or other characteristic of a product he is indirectly helping to improve the product in terms of his own wishes and needs. It was common during World War II, for example, when surveys were being conducted on government programs, for a respondent to preface his responses

representative of an agency which is able to bring about change. For example, facts about personal income, often considered difficult to obtain in interviews, are more readily available to the researcher if the respondent believes that the information he proffers will help responsible people in the government develop policies which will contribute to the public welfare and to the well being of the respondent.

A second major type of motivation depends more directly upon the personal relationship between the interviewer and the respondent. It can be defined as follows. An individual is motivated to communicate with another when he receives gratification from the communication process and the personal relationship. Such motivation sometimes occurs because the interview offers the respondent an opportunity to talk about topics in which he is interested but which usually do not obtain adequate expression. This does not imply that the respondent in a research interview ordinarily obtains cathartic release (although this may be present at times). It does mean, however, that the respondent obtains satisfaction from talking with a receptive, understanding interviewer about something in which he is interested and in which he is involved. The reader will recognize this as one of the basic motives of patients in the psychotherapeutic interview.

Interviewers are often surprised to encounter this motivation in a research interview, in which the possibilities (or desirabilities) of a therapeutic type of relationship appear remote. Experience shows, however, that if the research interview is conducted properly, this motivation is often present. The relationship in many ways resembles the counseling relationship. Counselors and therapists have found that freedom of communication (even the communication of deep personality conflicts) is possible under the proper conditions. Rogers, for instance, identifies four qualities which he claims are characteristic of the productive counseling atmosphere (18). Three of these four are relevant to the research interview. He characterizes the qualities as, first, "a warmth and responsiveness" on the part of the counselor which "expresses itself in a genuine interest in the client and an acceptance of him as a person." The second quality is described as "permissiveness in regard to expression of feeling. By the counselor's acceptance of his statements, by the complete lack of any moralistic or judgmental attitude, by the under

done by educating him in, or clarifying for him, the type of response expected, thus demonstrating to him his ability to handle the prescribed role. If the initial phase of the interview is successful, the interview has reached the point where one of two major types of motivation may be tapped thereby ensuring continued cooperation by the respondent.

Recent work in communications within small groups has resulted in findings which appear relevant to our understanding of communication between an interviewer and a respondent. These findings can be summarized as follows. One of the motives for communicating is the desire to influence, in some manner, the person to whom the communication is addressed. That is, a person will communicate in a given situation if he believes that such communication will bring about a change or effect an action which he considers desirable (9). In the interview situation, this means either that the interviewer will be perceived as a person who can bring about change directly or that the interview will be seen as a potential vehicle for indirectly accomplishing a desired change. A clinical psychologist, a social worker, or physician is frequently perceived by the patient or client as a direct agent of change. The respondent is likely to feel that communicating his symptoms or his financial difficulties to the professional person will benefit him directly. An example of a less direct relationship between interviewer and the goal of a respondent is provided by the typical market research survey, in which the respondent believes that by his expressed preference for a specific kind of packaging or other characteristic of a product he is indirectly helping to improve the product in terms of his own wishes and needs. It was common during World War II, for example, when surveys were being conducted on government programs for a respondent to preface his responses with "You tell those people in Washington that I said . . ."

This type of motivation can be developed only if the following relationships are apparent to the respondent. (1) The perception of the content of the interview as relevant to a change which he desires. The respondent will not spontaneously perceive every research project to be related to his goals and interests. The researcher must demonstrate this relationship or suffer the results of reduced respondent motivation to communicate. (2) The perception of the interviewer as a person who can bring about change or as the

representative of an agency which is able to bring about change. For example, facts about personal income, often considered difficult to obtain in interviews, are more readily available to the researcher if the respondent believes that the information he proffers will help responsible people in the government develop policies which will contribute to the public welfare and to the well being of the respondent.

A second major type of motivation depends more directly upon the personal relationship between the interviewer and the respondent. It can be defined as follows. An individual is motivated to communicate with another when he receives gratification from the communication process and the personal relationship. Such motivation sometimes occurs because the interview offers the respondent an opportunity to talk about topics in which he is interested but which usually do not obtain adequate expression. This does not imply that the respondent in a research interview ordinarily obtains cathartic release (although this may be present at times). It does mean, however, that the respondent obtains satisfaction from talking with a receptive, understanding interviewer about something in which he is interested and in which he is involved. The reader will recognize this as one of the basic motives of patients in the psychotherapeutic interview.

Interviewers are often surprised to encounter this motivation in a research interview, in which the possibilities (or desirabilities) of a therapeutic type of relationship appear remote. Experience shows, however, that if the research interview is conducted properly this motivation is often present. The relationship in many ways resembles the counseling relationship. Counselors and therapists have found that freedom of communication (even the communication of deep personality conflicts) is possible under the proper conditions. Rogers, for instance, identifies four qualities which he claims are characteristic of the productive counseling atmosphere (18). Three of these four are relevant to the research interview. He characterizes the qualities as, first, a warmth and responsiveness on the part of the counselor which expresses itself in a genuine interest in the client and an acceptance of him as a person. The second quality is described as "permissiveness in regard to expression of feeling. By the counselor's acceptance of his statements, by the complete lack of any moralistic or judgmental attitude, by the under

done by educating him in, or clarifying for him, the type of response expected, thus demonstrating to him his ability to handle the prescribed role. If the initial phase of the interview is successful, the interview has reached the point where one of two major types of motivation may be tapped, thereby ensuring continued cooperation by the respondent.

Recent work in communications within small groups has resulted in findings which appear relevant to our understanding of communication between an interviewer and a respondent. These findings can be summarized as follows. One of the motives for communicating is the desire to influence, in some manner, the person to whom the communication is addressed. That is, a person will communicate in a given situation if he believes that such communication will bring about a change or effect an action which he considers desirable (9). In the interview situation, this means either that the interviewer will be perceived as a person who can bring about change directly or that the interview will be seen as a potential vehicle for indirectly accomplishing a desired change. A clinical psychologist, a social worker, or physician is frequently perceived by the patient or client as a direct agent of change. The respondent is likely to feel that communicating his symptoms or his financial difficulties to the professional person will benefit him directly. An example of a less direct relationship between interviewer and the goal of a respondent is provided by the typical market research survey, in which the respondent believes that by his expressed preference for a specific kind of packaging or other characteristic of a product he is indirectly helping to improve the product in terms of his own wishes and needs. It was common during World War II, for example, when surveys were being conducted on government programs, for a respondent to preface his responses with "You tell those people in Washington that I said . . ."

This type of motivation can be developed only if the following relationships are apparent to the respondent: (1) The perception of the content of the interview as relevant to a change which he desires. The respondent will not spontaneously perceive every research project to be related to his goals and interests. The researcher must demonstrate this relationship or suffer the results of reduced respondent motivation to communicate. (2) The perception of the interviewer as a person who can bring about change or as the

representative of an agency which is able to bring about change. For example, facts about personal income, often considered difficult to obtain in interviews, are more readily available to the researcher if the respondent believes that the information he proffers will help responsible people in the government develop policies which will contribute to the public welfare and to the well-being of the respondent.

A second major type of motivation depends more directly upon the personal relationship between the interviewer and the respondent. It can be defined as follows: *An individual is motivated to communicate with another when he receives gratification from the communication process and the personal relationship.* Such motivation sometimes occurs because the interview offers the respondent an opportunity to talk about topics in which he is interested but which usually do not obtain adequate expression. This does not imply that the respondent in a research interview ordinarily obtains cathartic release (although this may be present at times). It does mean, however, that the respondent obtains satisfaction from talking with a receptive, understanding interviewer about something in which he is interested and in which he is involved. The reader will recognize this as one of the basic motives of patients in the psychotherapeutic interview.

Interviewers are often surprised to encounter this motivation in a research interview, in which the possibilities (or desirabilities) of a therapeutic type of relationship appear remote. Experience shows, however, that if the research interview is conducted properly, this motivation is often present. The relationship in many ways resembles the counseling relationship. Counselors and therapists have found that freedom of communication (even the communication of deep personality conflicts) is possible under the proper conditions. Rogers, for instance, identifies four qualities which he claims are characteristic of the productive counseling atmosphere (18). Three of these four are relevant to the research interview. He characterizes the qualities as, first, "a warmth and responsiveness" on the part of the counselor which "expresses itself in a genuine interest in the client and an acceptance of him as a person." The second quality is described as "permissiveness in regard to expression of feeling. By the counselor's acceptance of his statements, by the complete lack of any moralistic or judgmental attitude, by the under-

standing attitude which pervades the counseling interview the client comes to recognize that all feelings and attitudes may be expressed. No attitude is too aggressive, no feeling too guilty or shameful to bring into the relationship. A third characteristic of the productive counseling relationship is freedom from any type of pressure or coercion. The skillful counselor refrains from intruding his own wishes, his own reactions or biases into the therapeutic situation.

Applying these criteria to the research interview we may conclude that optimum communication takes place if the respondent perceives the interviewer as one who is likely to understand and accept his basic situation. The interviewer is thereby perceived as being within range—that is, he is seen as a person who will accept the respondent's statements and experience. This does not mean that the respondent must see the interviewer as similar to himself, but he should view the interviewer as capable of understanding or of being completely tolerant of his point of view. This perception will depend far more on the interviewer's attitude and the sort of relationship which the interviewer establishes with the respondent than on such external factors as the interviewer's dress or appearance, even though these externals may provide some initial cues for the respondent.

Unfortunately, a number of factors may inhibit communication or distort the content of the information given by the respondent. For example, the respondent may not accept the goal which the interviewer describes as the purpose of the survey. Thus, the goal of providing the federal government with information on income distribution may not appear worth while to some respondents when they are questioned about their own finances and economic situation. Even more frequently, the respondent may possess goals which are in conflict with the purpose of the interview. In an industrial plant, for example, a worker may be wholly sympathetic to the notion that employee opinions should be solicited and he may be hopeful that frank expressions may result in improving certain situations. However, he may also think that his expression of critical opinions may be dangerous and may make him liable to retaliation or discrimination. He may even be concerned over loss of work or *loss of promotion*. Such an attitude does not result in a refusal of an interview, but it does limit the content areas about which the

worker will talk freely. He might discuss physical working conditions, wage policy, and the like quite frankly but be extremely guarded and reticent when discussing any aspects of his job which might be interpreted as criticism of his immediate superiors.

Just as a respondent may refuse to communicate or may distort his communication because he rejects the research goals of the interview process, he may also refuse or restrict communication because the personal relationship with the interviewer is such that real communication and understanding is impossible. This kind of *respondent reaction may occur as a result of a stereotyped judgment* which he makes of the interviewer. Thus, if the respondent perceives a gap of education or an economic difference between himself and the interviewer, he may decide that the interviewer is incapable of understanding the respondent's family circumstances or of empathizing in any way with his predicament. This problem in communication is very likely to arise when the respondents have some perception of themselves as deviant in the content area of the interview. For example, a respondent holding extremely radical political views might perceive the interviewer as necessarily differing so greatly from himself that no tolerance for his point of view is possible. In such instances, the respondent has in effect concluded that the interviewer is outside the possible range of communication on the topic in question. The possibility of a complete and valid interview is, therefore, remote.

The relationship between the interviewer and the respondent and the character of the information communicated is not, however, determined exclusively by the respondent's stereotypes. There are a number of studies which document the importance of the interviewer's attitudes and biases in determining the results of the interview. Most of these studies fail to indicate specifically the manner in which the interviewer bias affects interview results. It seems safe to assume, however, that, although many of these situations may reflect the interviewer's failure to use the prescribed techniques, some are caused by a failure of interpersonal relations between the interviewer and respondent. Just as the respondent may behave according to his stereotyped judgments about the interviewer, the interviewer may be guided by his own stereotypes rather than by the objective characteristics of the respondent. The interviewer may.

for example, set up inaccurate working hypotheses as to how the respondent will respond and then, without awareness, guide or distort the responses into the anticipated channels (10)

In the two following sections of this chapter, we shall discuss those principles of questionnaire construction and interviewing technique which in the light of present experience are most likely to maximize a respondent's motivation to communicate and which help the interviewer to avoid the kinds of inhibiting or distorting factors which have just been described

DESIGN OF THE QUESTIONNAIRE

Dual Purpose of the Questionnaire

The questionnaire, or interview schedule, serves two major purposes. First, it must translate the research objectives into specific questions, the answers to which will provide the data necessary to test the hypotheses or explore the area set by the research objectives. In order to achieve this purpose, each question must convey to the respondent the idea or group of ideas required by the research objectives and each question must obtain a response which can be analyzed so that the results fulfill the research objective. Moreover the question must perform these two functions with minimal distortion of the response which it elicits. That is in asking a question of the respondent, we assume that the respondent possesses an attitude or opinion, or piece of knowledge. Each question should, therefore, be constructed so as to elicit a response which accurately and completely reflects each respondent's position.

The second function of the questionnaire is to assist the interviewer in motivating the respondent to communicate the required information. There are many factors which determine the respondent's willingness to engage in an interview, as we have already mentioned. In motivating the respondent, the skills of the interviewer are of great importance, of course, but the questionnaire itself does much to determine the character of the interviewer respondent relationship and, consequently, the quantity and quality of the data collected.

Since the questionnaire is constructed on the basis of the re

search objectives, it is clear that constructing the questionnaire cannot be the first step in undertaking a research project. The statement of the research objectives and the specification of the data required to meet those objectives must precede questionnaire construction. The sequence of steps in planning a research project is discussed in some detail in Chapters 1 and 2. It will be sufficient here to provide an example of the process by which a research hypothesis determines questionnaire content.

Suppose that, as part of a study of how behavior is influenced by mass persuasion, we have the hypothesis that the number of government savings bonds purchased is related directly to the amount of direct personal solicitation. What data are required to test this hypothesis? What questions should be asked to elicit these data?

In the present example, the investigators decided that two approaches should be employed, one direct and one indirect. The direct approach consisted of asking recent bond purchasers what factors had led them to buy. The indirect approach during a later portion of the same interview led to inclusion of a number of questions concerning the respondent's recent exposure to such influences as newspaper advertisements, radio, other group appeals, and individual solicitation. In analyzing the obtained data, the researchers sorted those respondents who were comparable with respect to income and other demographic characteristics into groups according to the frequency and type of solicitation which they had experienced. *The buying behavior of these groups was then studied, and it was found that buying behavior was closely related to the presence or absence of personal solicitation (7).*

In this example, one can see how the questionnaire design flows logically from the specified research objectives and must anticipate the analysis of the data. Thus, construction of the questionnaire is an integrated step in getting a research project into operation.

As noted above, the second function which the questionnaire must perform is to assist in creating conditions under which the respondent will communicate fully and freely. Research workers are by no means agreed on the techniques by which this can best be achieved (12). A systematic methodological research on the interviewer-respondent relationship has recently been completed by the National Opinion Research Center, and the final report is now in

the process of publication (10) The experience of the present writers argues that the characteristic of respondent orientation is of primary importance in maximizing communication The concept of respondent orientation as a characteristic of the questionnaire has an obvious analogy in the clinical concept of client centeredness (17) The comparison is more than a superficial one The questionnaire makes use of some of the techniques and to a small degree serves some of the purposes of client centered therapy The differences between the research and the therapeutic interview are many of course In the research interview content is determined primarily by research objectives rather than by the respondent's needs similarly the pace and sequence of questions is for the most part beyond the control of either the respondent or the interviewer Even the decision to engage in the interview is not wholly of the respondent's volition In spite of these limitations there are a number of respects in which the interview may be oriented toward the respondent and the questionnaire so constructed that the needs and reactions of the respondent are taken into account

The preceding pages have presented the major criteria by which a questionnaire might be judged However valid these criteria may be they do not solve the specific problems of question wording and question sequence which confront every social scientist who uses the interview The remainder of this section is devoted to a discussion of just such specifics a discussion which attempts to develop the do's and don't's of questionnaire construction out of the major purposes which the questionnaire must serve The topics included by no means exhaust the subject More detailed treatments have been made by Payne (16) Parten (15) Cantril (6) Blankenship (5) and others The problems discussed here are those which appeared most relevant in terms of the criteria cited and therefore most important in creating an adequate instrument for collecting social data by means of interviews

Language

In the construction of a questionnaire the primary criterion for the choice of language is that the vocabulary and syntax should offer maximum opportunity for complete and accurate communication of ideas between interviewer and respondent Not only should

the words chosen be within range of the respondent's vocabulary, but also his colloquialisms and clichés should either be known and used meaningfully or avoided. Most simply stated, the language of the questionnaire must approximate the language of the respondent. In most *cross section* surveys, one strives for simplicity, and properly so. For a study of physicians or lawyers, however, different and specialized vocabularies would be more appropriate. To the extent that the respondent population is not homogeneous, compromise is unavoidable. In such cases the solution of the problem consists in using language which communicates successfully to the least sophisticated of the respondent population and, at the same time, avoids the appearance of oversimplification.

Frame of Reference

To say that a questionnaire should be cast in the language of the respondent is relatively unequivocal and straightforward. It is equally important, and considerably more difficult, however, to phrase questions which take account of the frame of reference which respondents bring to the subject under discussion. Nevertheless, the questionnaire must introduce each topic in a form which ties into the perceptions of the respondent and is consistent with the respondent's notions of what is and is not salient to the topic under discussion. The development of a topic from one question to another must not only meet the researcher's criteria for reasonableness and logic, it must also meet those of the respondent. Thus, frame of reference becomes another dimension in which the researcher must begin at the point where the respondent is—it must be respondent oriented.

Bancroft and Welch (2) present an example of the effect of the respondent's frame of reference on his replies. They found that the series of questions used by the Bureau of the Census to ascertain the number of people in the labor market consistently underestimated the number of employed persons. When asked the question "Did you do any work for pay or profit last week?" respondents reported what they considered to be their major activity. Young people attending college considered themselves to be students, even if they were also employed on a part time basis. Women who cooked, cleaned house, and raised children spoke of themselves as house

the process of publication (10) The experience of the present writers argues that the characteristic of 'respondent orientation' is of primary importance in maximizing communication The concept of respondent orientation as a characteristic of the questionnaire has an obvious analogy in the clinical concept of client centeredness (17) The comparison is more than a superficial one The questionnaire makes use of some of the techniques and to a small degree serves some of the purposes of client centered therapy The differences between the research and the therapeutic interview are many, of course In the research interview, content is determined primarily by research objectives rather than by the respondent's needs similarly, the pace and sequence of questions is for the most part beyond the control of either the respondent or the interviewer Even the decision to engage in the interview is not wholly of the respondent's volition In spite of these limitations, there are a number of respects in which the interview may be oriented toward the respondent, and the questionnaire so constructed that the needs and reactions of the respondent are taken into account

The preceding pages have presented the major criteria by which a questionnaire might be judged However valid these criteria may be, they do not solve the specific problems of question wording and question sequence which confront every social scientist who uses the interview The remainder of this section is devoted to a discussion of just such specifics a discussion which attempts to develop the 'do's and don'ts' of questionnaire construction out of the major purposes which the questionnaire must serve The topics included by no means exhaust the subject More detailed treatments have been made by Payne (16), Parten (15) Cantril (6), Blankenship (5), and others The problems discussed here are those which appeared most relevant in terms of the criteria cited and, therefore, most important in creating an adequate instrument for collecting social data by means of interviews

Language

In the construction of a questionnaire, the primary criterion for the choice of language is that the vocabulary and syntax should offer maximum opportunity for complete and accurate communication of ideas between interviewer and respondent Not only should

a question, there is an implication that the respondent should be in possession of an adequate answer and that if he cannot answer, he is somehow discredited. If, for example, in a questionnaire dealing with public attitudes toward problems of atomic energy, an interviewer asks a respondent, 'What precautions are appropriate for a technician handling radioactive isotopes?' an immediate and very common respondent reaction will be embarrassment and resentment at being asked a question which he is unable to understand. Not only will the researcher have lost the answer to a question, but he will also pay a heavy price in terms of decreased motivation to communicate. Another possibility, of course, is that the respondent, feeling obligated to show his knowledge will pretend to competence which he does not possess.

The importance of asking questions appropriate to the respondent's level of information, and not productive of respondent embarrassment, does not necessarily limit us to asking questions to which every respondent knows the answer. It does mean, however, that caution in wording questions must be used when we anticipate that a considerable proportion of respondents will not be in possession of the answer. For example, the question quoted above might be preceded by a statement such as "Many people haven't had an opportunity to learn a great deal about the technical problems of handling atomic material, but some have picked up information on this subject. Do you happen to know . . . ?"

This problem is sometimes referred to as *expert error*—that is, the error of ascribing to the respondent a degree of expertness in a particular field which he does not actually possess. These 'expert' questions may require the respondent to engage in uncomfortable self analysis, to be verbal about material which is really unanalyzed or un verbalized and therefore not consciously available. Suppose we ask an industrial employee, 'What is the state of your morale, and what is the reason for your feeling that way?' It would be as if a doctor asked a patient for the name and cause of his disease, rather than asking for the patient's symptoms from which the nature of the disease may be inferred.

Social Acceptance

Another characteristic of the respondent centered questionnaire

wives even if they also did some work for pay outside the home. The effect of the respondent's frame of reference was to classify as nonworkers many thousands of people who met the census definition of workers. The solution involved revising the sequence of interview questions beginning with the acceptance of the respondent's classification of himself. Thus, people were asked first what their major activity was. Then, those people who gave 'nonworker' responses were asked whether, in addition to attending school or keeping house, they did any work for pay. The effect of this change was to raise the official estimate of employment by more than one million persons.

The respondent's frame of reference may also be important in determining whether he will be willing to communicate a given piece of information. He may be reluctant to communicate if he fails to see the relationship between a question and his own perception of the research objectives. Thus a survey respondent who has talked freely about foreign policy may suddenly balk at being asked his age or education. Although these are unlikely to be threatening questions, they do not fit his perceptions of the research needs. The collection of data on family income cited earlier in this chapter, provides another example of the extent to which respondent behavior may depend upon his perception of what is relevant. The collection of detailed data on personal income, unsuccessfully attempted in many early studies, was achieved by introducing a request for income data as part of a program to assess, and perhaps ultimately to solve problems of consumer credit, spending, and saving. In the context of discussing savings, plans for consumer purchases, and attitudes and expectations about economic and personal financial status, the question of family income appears to the respondent as reasonable and relevant.

Information Level

A question must be worded so that it ties into the respondent's present level of information in a meaningful way. No unrealistic assumptions should be made about the expertness of the respondent or the amount of information he possesses. The importance of this rule for questionnaire construction lies in the fact that when the interviewer, with the authority of his role, asks the respondent

merely agreeing with the language of the question. It is more difficult to respond, "I disagree," since this response seems to contradict the interviewer, or at least goes counter to the ideas of the person who worded the question.

One way in which a question may suggest a positive or negative answer is through the use of words which have become emotionally loaded," either favorably or unfavorably. In our culture there are many words so affect laden that it is virtually impossible to expect a respondent to give a response to the concept behind these words. For example, prior to World War II, 'Nazi' had already become an emotionally colored word. As a result, one obtained very different responses to a question referring to 'Nazi Germany,' rather than simply to "Germany."

Another way in which a question may encourage a particular response is by associating one of the alternative responses with a goal so desirable that it can scarcely be denied. Thus, the question

'Do you favor or oppose higher taxes to prepare for the dangers of war?' associates higher taxes with defense against attack and implies that a negative answer reflects indifference to the menace of attack. Even if the respondent is permitted an unstructured reply, the question does much to bias his answer. If he is given only the alternatives of acceptance or rejection, as in the typical public opinion poll, the biasing effect of the question is even more serious.

A loaded question is not necessarily undesirable and often has a real place in the questionnaire. The problem is to avoid loading if one is looking for an undistorted response. The following is an example of a strongly loaded question which was purposely used in one study: 'Would you favor sending food overseas to feed the starving people of India?' In this case the question followed a series of unloaded questions and was used to determine the number of people who were so strongly against shipping food that they rejected the idea *in spite of* the strong emotional context of 'starving people.'

The Single Idea

Questions should be limited to a single idea or to a single reference. The problems encountered in this area are illustrated in the following question: 'Do you favor or oppose unemployment

is its emphasis on the acceptability of a wide range of responses. No question should confront the respondent with the necessity of giving a socially *unacceptable* response. If we expect the respondent to answer freely and spontaneously we must help him to feel that the entire range of possible responses is acceptable—acceptable not only to the interviewer but also in terms of the respondent's own standards for himself. For example, if after a presidential election we wish to ascertain who in the population did and who did not vote, we find ourselves in the position of asking respondents about a situation in which they may perceive only one socially acceptable alternative. The civic-minded responsible citizen voted; therefore the respondent voted, or at least should have voted and does not wish to be put in the position of telling the interviewer that he failed to do so. This hesitancy can be overcome at least in part through question wording. For example, some form such as the following might be used: You know, in the last election about half of the qualified voters actually got to the polls and about half were unable to—did you happen to vote?

Offering a range of responses which meets the respondent's criteria of social acceptability is necessary to good question formulation. A broader statement might be that the question must never constitute a threat to the respondent's ego. Such a threat may be introduced if the respondent is required to give an answer which he feels is socially unacceptable, or it might come about if the respondent is placed in a position where he feels less well informed than he should be.

Leading Questions

Questions should be phrased so that they contain no suggestion as to the most appropriate response. For example, a question designed to elicit general attitudes toward rent control might read: How do you feel about rent control? A form of the same question which is obviously biased might be: You wouldn't say that you were in favor of rent control, would you? This kind of bias is so easily recognized that we avoid it almost without effort. A more subtle form of biased wording might be: Would you say that you are in favor of rent control? This question makes it easier for the respondent to answer yes than no. In answering yes, he is

sequence of questions the frame of reference is gradually narrowed by asking more specific questions. The purpose of the funnel sequence is to prevent early questions from conditioning those which come later and to ascertain from the first open questions something about the respondent's frame of reference. The following series of questions illustrates the funnel approach.

- QUESTION 1 How do you think this country is getting along in its relations with other countries?
- QUESTION 2 How do you think we are doing in our relations with Russia?
- QUESTION 3 Do you think we ought to be dealing with Russia differently than we are now?
- QUESTION 4 (If yes) What should we be doing differently?
- QUESTION 5 Some people say we should get tougher with Russia and others think we are too tough as it is. How do you feel about it?

The reader will notice that the first question is very general in its approach. It does not establish a frame of reference—a trend of thought—in regard to the country under discussion in terms of diplomatic relations or of economic relations. It permits the respondent great freedom in discussing the topic. From the answer to the first question, we can probably infer the frame of reference of the respondent. In the second question we have restricted the area to one country, Russia. The third question is aimed at the respondent's opinion of how the United States ought to deal with Russia, and the fifth becomes very specific by asking whether we should exert more pressure or be more lenient. If for example Question 5 had been asked any earlier in the sequence, it might well have conditioned the answers to the other questions. The funnel technique is, therefore, often very helpful in avoiding the distortion of a question by those which precede it. It enables us to analyze the frame of reference which the respondent is taking and it enables us to get a general affective response before pinning the person down to more specific points.

The first two or three questions in a questionnaire often have a dual function. On the one hand they are included to obtain information on specific research objectives, but they also help to educate and motivate the respondent. In many instances the re-

insurance and pension plans?" Many answers to this question would not permit the researcher to determine whether the respondent is answering one or both of the items mentioned in the question. The most acceptable formulation of a question in this area would depend to some extent on the specificity of the research objectives. If the purpose is to find out the respondent's attitude toward pensions and unemployment insurance specifically, it would be necessary to ask two questions, one referring to each of the two proposals. If, on the other hand, the purpose of the question is to get some notion of the respondent's general attitude in the area of worker benefits, it may be possible to ask a global question such as "Do you favor or oppose such worker benefits as unemployment insurance, pension plans, and the like?" It must be kept in mind, however, that if a global question of the latter type is asked, the interpretation must be very conservative. In other words, a positive response to a global question must be taken only to indicate favorableness in the general area and cannot be interpreted as indicating respondent support for any of the specific examples cited.

Question Sequence

Aside from the wording of individual questions, the researcher needs to give thought to the arrangement of the questions in a questionnaire. At several points in this chapter we have discussed the concept of respondent orientation, which is also important in considering question sequence. Thus, questions should be so arranged that they make the most sense to the respondent—that is, the sequence of ideas in a questionnaire should follow the logic of the respondent. It may be that questions which are associated together from the analyst's point of view are widely separated in the questionnaire. The sequence of questions should be determined primarily by the interview process rather than the research process. A well-designed questionnaire facilitates the easy progress of the respondent from item to item and often leads him to anticipate the next question because it seems to him the logical topic to discuss.

The sequence of questions may also be determined by what is called the "funnel approach." This refers to a procedure of asking the most general or the most unrestricted question first and following it with successively more restricted questions. Thus, in the

contained in the question so that the respondent merely has to select the category which comes closest to his position. An example of a closed question is, 'Do you think your income will be higher, lower, or about the same this year as it was last year?'

Generally speaking, the closed question is well adapted to situations in which (1) there is only one frame of reference from which the respondent can answer the question (2) within this single frame of reference, there is a known range of possible responses and (3) within this range, there are clearly defined choice points which accurately represent the position of each respondent. Two examples will help to clarify these points. The first is the classification of respondents by marital status. In this case, there is a known range of possible responses: a person is either single, married, divorced, separated, or widowed. Within this range, the choices are clear and the question has but one frame of reference for all respondents. Here the closed question is desirable and can be worded: 'Are you single, married, divorced, separated, or widowed?' The respondent merely has to select the response which defines his marital status.

Another example of the closed type is the question: 'Would you say your present income was higher, lower, or about the same as your last year's income?' In this question the respondent is asked to compare two facts which are known to him. The frame of reference is limited to an income comparison for two years, and the choices are clear.

Crutchfield and Gordon have provided an excellent documentation of the effects of using the closed question improperly (8). The following question was asked on a national survey: 'After the war would you like to see many changes or reforms made in the United States or would you rather have the country remain pretty much as it was before the war?' The answers indicated that the majority of people wanted things to remain as they were. A follow up was made in which the same question was asked and nondirective probes were then used to ascertain what the respondents were concerned with when answering the question. The responses showed that respondents answered from seven frames of reference. Some were concerned with domestic issues (employment conditions, standard of living, etc.), some with technical improvements (better transportation, communications, etc.), others with political affairs, and so on. Since the original researcher was unaware of the varying

respondent may not know the kind of responses expected of him—that is, whether a one-word answer will suffice or whether he is being asked to discuss the subject in detail. He may wonder about being cross-examined, or he may be simply confused about the interview demands regardless of the initial instructions. During the first two or three questions the interviewer, by his probing, his reactions to responses, and his general behavior toward the respondent, educates the respondent in the role which is expected of him in the interview.

In addition to their orienting or educational purpose, the first questions also serve to motivate the respondent to participate more thoroughly by involving him in the topic under consideration. The first few questions may, in fact, set the tone of the entire interview.

Often a questionnaire covers more than one general topic. This may create difficulties in the interview, since the respondent has to be helped to change his frame of reference from one topic to the next. One efficient way to help the respondent orient himself to a new area of discussion involves the use of transition statements or transition questions. Such a statement might be, "Well, we've been discussing our relations with the Far East; now we want to talk a bit about the way things are going for us in Europe. How do you feel about our relations with the countries in Europe?" Statements of this sort help the respondent to shift gears and to transfer his attention to a new area of discussion.

The Form of the Question

Thus far we have discussed the wording of questions without considering the problem of the form of the response—that is, whether the respondent is to reply in his own words or whether he is to select from a series of preassigned categories the response coming closest to his own opinion. Questions of the former type are termed "open" or "unrestricted"; the latter type of question is "restricted" or "closed." The open question is one in which the topic is structured for the respondent but he is given the task of answering in his own words, structuring his answer as he sees fit, and speaking at whatever length he desires. An example of an open question is, "How do you feel about Negroes and whites working together in this factory?" In the closed question the possible responses are

encountered in coding responses. In both types the respondent's attitude or perception must be categorized. If the question is closed, the categorization is made by the respondent or the interviewer. If the question is open, the categorization may be made by the interviewer, it is usually considered preferable, however, to have the responses coded at some central place by people trained for this work. Each of these coding procedures has advantages and disadvantages, which are discussed and evaluated in Chapter 10. For a more detailed discussion of open and closed questions and their respective uses, the reader should see Lazarsfeld (13). Lazarsfeld contends that the methods can be combined effectively if the open question is used in an elaborate pretest followed by closed questions in the main study with open questions used in a follow up of critical cases.

The Pretest

No matter how astute the researcher has been in wording his questions and designing his questionnaire, he needs to try them out with respondents before launching into the actual field studies. The pretest is, in a sense, a miniature study in itself. The first function of the pretest is aimed at testing the questionnaire from the research point of view. The interviews should be analyzed to see whether the responses fulfill the research objectives. Some of the researcher's "best questions" often fail to elicit the type of response which meets the objectives. An analysis of these trial interviews in relation to the objectives will increase the probability of fulfilling the research objectives. Often the pretest calls for major revision of the questions, and several pretests are required until a workable questionnaire is achieved.

A second objective of the pretest is to determine the extent to which the questionnaire meets the criterion of respondent orientation in all its aspects. Does the questionnaire promote the appropriate relationship with respondents? Do respondents understand the questions? Can the questions be asked without having to be explained or reworded? There are no exact tests for these characteristics. The help of experienced interviewers is most useful at this point in obtaining subjective evaluations of the questionnaire.

frames of reference, his interpretation of the findings was quite in error. This illustrates the danger of using closed questions when more than one frame of reference is possible.

Let us consider another example. At the present time, do you think personal income taxes are too high, too low, or about right? The alternatives within the question are incomplete and fail to allow for the person who feels that the income taxes are fair at certain income levels but are unfair at other levels. Neither are the alternatives adequate for the person who feels that taxes are too high because of heavy government expenditures but that if the expenditures continue then taxes must remain high. Thus, for many people the alternatives presented do not include choice points which closely approximate their attitudes. Such people are forced either into discussing the topic more fully with the interviewer or into selecting a category which is a poor approximation of their position.

The open question has many advantages stemming from the fact that the respondent is encouraged to structure his answer as he wishes. The technique provides a means of obtaining information which cannot be obtained adequately by use of a closed question. For example, it permits the respondent to state his own frame of reference when this is desirable. The potentialities of the open question for discovering motivation have been ingeniously explored by Lazarsfeld (14).

Another advantage of the open question is the information which the answers indicate with respect to the respondent's level of knowledge or degree of expertness. If the respondent has been led to discuss his opinions of the Atlantic Charter, one is able to analyze not only his attitude but also his level of information.

The relatively free interchange between interviewer and respondent which is characteristic of the open question permits the interviewer to discover whether the respondent clearly understands the question which is being asked of him. On the other hand, once the respondent has selected one of the proffered alternatives to a closed question, the interviewer can assume only that the respondent understood the question and chose the alternative which best approximated his own position.

Another difference between the open and closed question is

situation only in the absence of certain specific kinds of barriers to communication

When the interviewer first faces a respondent, he finds that the relationship has some structure even before a word is spoken. On the one hand, the respondent probably will be polite enough to let him talk, on the other hand, there may already exist certain barriers which must be overcome. For example, the public opinion interviewer is frequently mistaken for a salesman by the respondent. Another barrier to communication arises from the respondent's perception that granting the interview will in some way make him vulnerable. This is essentially a problem in reassuring the respondent with respect to the anonymous or confidential nature of the interview. A third barrier comes from a rather frequent respondent perception that the interview may be intended in some subtle way to check up on him or his activities. Handling this sort of problem calls for a convincing explanation by the interviewer of the purpose of the study and particularly of the method by which the respondent was selected.

The positive motivation in terms of the respondent's goals comes from a careful statement of the purpose of the research. The interviewer tries to sense the respondent's wishes or goals with respect to the interview process and having appraised these to explain to the respondent how the interview relates to them. For example, an interviewer working on a study of public attitudes toward current matters of foreign policy might come upon a respondent who on hearing the purpose of the survey, says to the interviewer: "You don't want to talk to me about foreign policy. What I think about those fellows in the State Department would curl their hair. You'd better find somebody who is a more agreeable type." The interviewer would assure the respondent that the purpose of this study is not to find out simply the opinions of people who are endorsing current foreign policies. He would emphasize that the interview provides an opportunity for the respondent to register his criticism in a place where it might have some effect on public officials who sincerely want to learn the general public attitude, whether it is critical or appreciative.

In some researches the respondent goal is rather clearly perceived, such as in the case of the worker who is asked to participate in a study which may result in better working conditions or higher

PRINCIPLES OF INTERVIEWING

The preceding section dealt with the instruments of data collection. This section will discuss the specific techniques which the research interviewer uses. The techniques proposed are a systematic, well tested set of procedures which are consistent with the principles of communication discussed earlier in this chapter.

The Introduction to the Interview

The first step is often the most difficult for the interviewer, because at the initial contact the respondent must be motivated to permit the interview. Ordinarily the interviewer will follow a sequence of procedures approximately as follows:

- 1 Explain the purpose and objectives of the research
- 2 Describe the method by which the respondent was selected
- 3 Identify the sponsor or the agency conducting the research
- 4 State the anonymous or confidential nature of the interview

In the early phases of the interview, the interviewer plays one of his most important and one of his most autonomous roles. It is difficult to describe the precise acts which an interviewer should perform in order to provide adequate motivational basis for the respondent to communicate the information which he seeks. The establishing of rapport is clearly not a scientific procedure in the sense of being capable of objective statement. It is rather a skill which depends primarily on the know-how, experience, and sensitivity of the interviewer. It is this function of the interviewer which makes great demands on the qualities of clinical insight and intuition.

We have already mentioned that the forces leading a respondent to communicate can be thought of in terms of a means-end or path-goal sequence in which the respondent gives information because he sees the information giving process as a means of attaining some goal which he considers desirable. Secondly, the respondent is motivated to give accurate and complete information as a means of attaining some satisfaction out of the relationship with the interviewer. Thirdly, the respondent communicates in the interview

distinguish among interviewing situations according to the amount" of rapport which they require. Thus, it would be possible for an interviewer to do a very acceptable job of asking the two or three simple demographic questions associated with the typical school census without having established any relationship with the respondent beyond that implied in a civil "Good morning" and a display of credentials. On the other hand, if the interviewer's task is to obtain information about some aspect of the respondent's habits—marital relations, for example—it would be necessary for him to establish a deeper kind of personal relationship with the respondent. In general, we can say that the more intimate, emotionally charged, or ego involved the topic of the interview the more delicate the job of establishing the relationship with the respondent becomes, and the deeper that personal relationship must be.

When we refer to a deeper or a closer personal relationship as appropriate to certain kinds of interviewing, the things we have in mind are those associated with such words as warmth, acceptance, understanding, tolerance, and the like. We are not suggesting that the interviewer can do a more effective job if he is closely involved with the respondent's activities. Thus, for example, we are not suggesting that a close friend or near neighbor is the ideal interviewer, on the contrary, the ideal interviewer-respondent relationship seems to be one in which the interviewer achieves a considerable degree of closeness in terms of understanding and acceptance but at the same time retains the detachment or objectivity which we associate with a professional client relationship.

Asking the Questions

The interviewer's job of asking questions from the questionnaire is comparable to the scientific technician's role of applying a measuring instrument in a standard manner. It is through the use of carefully worded questions transmitted to the respondent verbatim that we achieve much of the standardization in the interview.

The major aim in putting questions to a variety of respondents is to have those questions so worded that their psychological value is equivalent for each respondent. There are infinite differences among respondents and it is not possible to vary a question so that

pay In other studies the respondent goal is more obscure In a laboratory research problem for example, the respondent may gain only the prestige of participating in a scientific or official endeavor

Another motivation the interviewer should tap comes from the personal relationship which he builds with the respondent In part, this relationship depends upon the interviewer's being perceived as a desired agent of communication or change However, as in the therapeutic interview the qualities of acceptance understanding and receptivity seem to have inherent values for the respondent Some evidence for the importance of the interviewer respondent relationship was obtained by the Survey Research Center when respondents who had been interviewed about their incomes, savings and buying plans were polled by mail about their reactions to the interview Their replies were more often couched in terms of the personal relationship and personal qualities of the interviewer than in terms of the content of the study or the apparent purpose of the inquiry Typical comments mentioned the fact that the interviewer was a very understanding person or that the interviewer had a keen insight into the respondent's situation

Often the interviewers' contribution to respondent motivation is referred to as rapport The term has come into common use and indicates an increasing sensitivity on the part of researchers to the importance of the interviewer respondent relationship At times the use of the term suggests a superficial approach to respondent motivation Thus rapport is referred to as if it were some tangible quantity or some specific task which was to be gotten out of the way early in the interview as a preamble to getting on with the main business of data collection There is the implication that after the interviewer has said Good morning and inquired after the health of the respondent's family with the properly solicitous inflection he can ignore the relationship with the person giving him data Contrary to the implications of this approach rapport is not something which is plugged in at the beginning of the interview in order to get it off to a good start Rapport refers to the atmosphere or climate of the entire relationship between respondent and interviewer

Although rapport or the climate of the interviewer respondent relationship has yet to be reduced to quantifiable factors we can

following types of situations (1) to elicit additional information from the respondent when further information is necessary to the research objective, and (2) to clarify or make more specific information which the respondent has already given. All of this must be accomplished without changing or biasing the data.

The "probing" techniques useful for such purposes can be classified generally as "nondirective." They enable the interviewer to act as a catalyst—that is, to bring about a reaction without himself becoming part of the reaction. The effect of such probing is to increase the intensity or "response getting" power of the stimulus question without changing its content or structure.

Thus, to gain more information the interviewer uses such phrases as "Would you tell me some more about that?" "I'm interested in what you're saying. Could you give me a little more about that?" or "I see what you mean. Can you tell me a little bit more about how you feel there?" These statements indicate that the interviewer is interested, understands what the respondent is saying, and is making a direct bid for more information. To accomplish the second task, that of clarification of information already given, the interviewer might use such probes as "Now let me see if I have it straight. As I get it, you feel . . ." and then summarize what the respondent has said. Or he might say, "I would like to read my notes back to you to see whether I have your point of view straight."

It is through the use of such nondirective probes that the interviewer does much to develop the permissiveness and warmth which are so important in the interview. The reader who is familiar with the literature on client-centered counseling will recall the stress which is placed on the atmosphere of permissiveness as the basis for permitting the client to examine his own attitudes. Such an atmosphere permits the client to verbalize the deeper attitudes which are usually concealed from outsiders. Many of the same dynamics are present at a more superficial level in the research interview, whether one is dealing with personal attitudes or factual data. Let us consider some examples of the effect of skillful probing.

I: How do you feel about sending money and help to other countries?

R: Well, I don't know. Sometimes I think we go too far.

I: I see. Can you tell me a little more about what you have in mind?

it has the same psychological impact for each. Since, therefore, we cannot tailor the question for each respondent, the best approximation to a standard stimulus is to word the question at a level which is understandable to all respondents and then to ask the question of each respondent in identical fashion. This, then, is the function of the interviewer in using the questionnaire as the stimulus. The only instance in which the interviewer is permitted to vary this procedure is when an individual is unable to understand the question as worded. Even in such cases, the interviewer is encouraged to repeat the question verbatim before explaining it. In many cases, apparent lack of understanding is a matter of attention fluctuation rather than inability to grasp the question meaning. In such instances, a simple repetition of the question will suffice.

Except for these minor variations, the interviewer's role with respect to the questionnaire is to treat it as a scientific instrument designed to administer a constant stimulus to a population of respondents. This technique is necessary when quantifiable data are desired. In some research of an exploratory nature, or where subjective analysis is contemplated, the interviewer may be permitted much more leeway in the use of the questionnaire. In some research he may tailor his questions to each respondent, with the researcher indicating only the areas to be investigated. Where quantifiable data are needed, however, the more rigid use of the questionnaire appears necessary.

Stimulating Complete Responses

In many cases the use of the question evokes a response which is incomplete or which is unclear. The interviewer must have some technique which will enable him to stimulate the respondent to further verbalization. Moreover, he must achieve this without sacrificing standardization. For example, if a question is asked of all respondents, we have, so far, comparability. If at this point each interviewer asks a different subquestion which he makes spontaneously, the responses are no longer responses to the original question but will vary from interviewer to interviewer depending upon the subquestion which was asked. This defeats the objective of standardization.

Specifically, the interviewer needs techniques to handle the

having served to make the point was dropped. The final answer of 400 bushels was almost certainly closer to the actual fact.

As indicated earlier in this chapter the general effect of this type of interpersonal relationship is pleasurable for the respondent—in that he has the opportunity to talk with a skillful interviewer. He reacts to the permissive, accepting atmosphere by communicating willingly with the interviewer.

Recording the Responses

One final job remains for the interviewer: this is to get a faithful and accurate report of the responses. Experience has shown that the only accurate way to reproduce the responses is to record them during the time of the interview either by mechanical methods or by having the interviewer take extensive notes. A good deal of relevant information is almost certain to be lost if the recording is left until the interview has been completed. It is not within the scope of this chapter to discuss various kinds of recording devices. Whatever the method, however, the interviewer must be trained in its use and it must be carried out faithfully during the process of the interview itself.

SAMPLE INTERVIEW

In order to demonstrate some of the techniques which have been discussed in this chapter, we have included a brief sample of a data collecting interview taken in an industrial plant manufacturing tractors. The respondent was a foreman. The example is an excerpt from a phonographically recorded interview. It has been edited slightly in places to make it more intelligible. The interviewer's questions preceded by an asterisk are questionnaire items. All others are the interviewer's probe. This example is selected not as an ideal interview but merely to demonstrate the techniques used by one experienced interviewer.

1 I *What do you do on your job?

1 The objective is to obtain a general picture of the type of work and responsibilities

- R Well maybe we ought to give some help but my gosh when I see our tax money go to help some of those countries who aren't doing much for themselves I think sometimes we'd better lay off
- I Sometimes you feel we ought not to help them
- R That's right! I think we'd better let them go their own way and to hell with them!

In this example, the respondent made a mildly critical statement at first. The interviewer reacted to this by being nonevaluative and yet accepting. He didn't criticize the respondent nor did he agree with him; he merely indicated a general acceptance of the statement. The result of this was a somewhat more pointedly critical statement. The nonevaluative acceptance by the interviewer permitted the respondent to make his final bitter response without feeling the need to defend or modify it.

The next example is taken from an interview with a farmer. The interview was concerned almost exclusively with problems of farm production.

- I How many bushels of wheat did you harvest last year?
- R My gosh we had a *terrible* year! When we'd ought to be planting last spring it rained all the time and then it got dry and everything burned up. We didn't get more than 300 bushel!
- I I see. Well you said you didn't get more than 300 bushels. Can you give me a little closer estimate?
- R Well like I said it was an awful year around here but I guess we got a little more than 300—between 350 and 400 I guess actually.
- I 350 to 400 you say. Which would be closest?
- R Oh I think we estimated it at right around 400 bushels.

Notice that the interviewer again began with a nonevaluative statement essentially repeating the respondent's first estimate. The effect was that the respondent revised his estimate of wheat yield. It seems apparent that his first response was more concerned with the misfortunes of the crop than with a precise estimate of it. The interviewer ignored the attitude and focused on the factual part of the response. The result was that the estimate of 300 bushels

formation already obtained and then a virtual verbatim repeat of the original question

R: Well, the track comes down the assembly line and the tractor comes down the assembly line and the last thing that we do to it is to put these steel tracks on it so that it can be driven away and, uh, we've got an electric hoist that lifts the heavy tracks and puts 'em in; we jockey 'em into place and then, uh, some of the men work on the top of the track fastening cleats to it, and some of the men work on the bottom and, uh, I kind of look to see that they're doin' it all right and help 'em out if there's any trouble (*Pause*)—keep track of our production

5. I: So, one of your jobs as foreman is to see that the men are doing their work properly as you indicate and, uh, keep track of production.

5. This illustrates the use of a content summary as a probing technique. The interviewer merely summed up the statements which were made. This device is particularly effective after a rambling, incoherent statement. The summary serves to focus attention on the central content of what has been said. In addition, it indicates to the respondent that he has been communicating ideas and that the interviewer has accepted the ideas. Usually, the summary stimulates further responses, either additional data or clarification of what has been reported previously.

R: Yes, that's right (*Pause*) and then every day they send me a report of what our work was like the day before, how much work we got out and how much scrap there was, and uh—it's up to me to see that the amount of work is okay and that there is not too much scrap.

The question was asked word for word as it appeared on the questionnaire

R Well, I'm a track foreman, that is, I'm in charge of the men who are putting together these metal tracks, you know, that tractors run on

2 I The tractors run on, you say?

2 *Unfortunately a transcription does not show inflections and emphasis In this case the interviewer's question had a slight rising inflection on the end, indicating a mild "I don't quite get you" probe This probe does not really lead in the direction of the objective of the question, but it does give the interviewer a better background for further answers by giving him information about the respondent's work*

R Yes, these big heavy tractors run on a steel track like a tank and, uh, after the tractors been assembled, we've gotta hang one of those heavy steel tracks on each side of them

3 I I see

3 *These brief, permissive, encouraging comments appear frequently throughout the interview This type of comment and nodding of the head, indicating and encouraging comments, are the most frequent "technique" which the interviewer uses In this recording many are lost in the reproduction*

R I'm in charge of the crew that does that

4 I Well, will you tell me a little bit more about your job—you say you're in charge of the crew, just what kind of things do you do?

4 *This serves to bring the respondent back into the area of the question objective The reader will note the brief summary of the pertinent in*

- 11 *I* I see Well, that gives me an idea of what your job is and how long you've been on the job *Now tell me, how do you feel about the job you have now?

11 This illustrates a brief transition state ment The interviewer by his remarks indicates that one area of the questionnaire is completed and another is to be introduced This is a useful technique to help the respondent shift his frame of reference to the new topic

R Well, it's better than when I was on the track line

- 12 *I* How's that?

12 Another probe which is used when clarification is wanted The inflection indicates 'I'm not quite sure I understand you Will you amplify that remark?'

R Well, for one thing, foreman's pay is higher and also it's more regular—it goes right on, it's not hourly

- 13 *I* Uh hum

R And, uh, besides that, I like the kind of work better

- 14 *I* Uh hum—okay, you say you like the job better than the one you had before that, uh, let's take it all in all, how do you feel about your job?

14 Up to this point the respondent has been answering in terms of details on his job The objective calls for general affect The interviewer tries to communicate this frame of reference in this way of re asking the question stressing the over all aspects

R Well, I guess I like it all right—it's got its headaches like all good jobs do, I guess

- 15 *I* Well, that's one thing we want to talk about, uh, you've given me some information

- 6 *I* I see, uh, are there any other kinds of things you do on your job?

6 *This is a very direct bid for additional information. The stress on the word "kinds" is directive in that it asks the respondent to shift his frame of reference.*

R Oh, I, uh, I take care of things like time off and, uh, figuring out, uh, uh, when the men can go on vacation without busting up the work schedule, and, uh, if, uh, a man has been around for a while and is about ready for a pay raise, uh, higher pay rate on the job he's doing, it's up to me to recommend em, sometimes if they want to promote a man to a higher job, uh, that's on my recommendation.

- 7 *I* Uh hum, uh—*How long have you been on this job?

R Oh a coupla years

- 8 *I* You've been on this job a couple of years?

8 *This restatement of the response brings additional specificity from the respondent.*

R Well, not quite so long—uh—let's see, I came on this job after Joe left, and that was a year ago Christmas time—it's about a year and a half, really.

- 9 *I* Year and a half, I see. *What were you doing before that?

R I was laying track.

- 10 *I* That was for the same company, you mean?

10 *Part of the objective of the question in 9 is to determine whether the person's prior job was in the same company. Here the interviewer uses a direct question to ascertain this information.*

R Yup, right here.

his attitudes on the foreman's role He broke into the discussion and returned to a topic which the respondent had mentioned earlier This was unfortunate, since he lost material which might have been very relevant on working with the men

R Well, what I mean is working with people—now, uh, I remember some of the things that used to seem good or bad to me when I was on the track line, and this way being foreman, I get a chance to try to make the set up a little better for the rest of the guys

- 19 I You have a chance to help the men some

19 This is a statement rather than a question It serves to summarize responses and encourage further responses

R Yeah, I remember how it was when I was on the line, and I think I can make things better

- 20 I Such as

20 This probe is the same as "What were you thinking of here?"

R Well, like making it handy to get tools they need and arranging the work so they don't have to work hard sometimes and loaf others things like that

- 21 I I see Well now, let's take the other side of the picture for a minute, uh—*What are some of the things that you don't like about your job?

R Oh, I don't really know what to say to that

- 22 I Uh hum

R I, uh, I don't like to complain you know, they've been pretty good to me here

on things you like about the job, but I was going to ask you, uh—*What are some of the things you like best about your job you have now?

15 One of the problems of an interviewer is how to ask a question when the respondent has given a partial answer under other questions. This illustrates a technique for handling the problem. The interviewer recognized that the respondent had mentioned something on the topic earlier, then he proceeded to ask the question. This avoids the implication that the interviewer was not attending to the earlier discussion and serves to get new material.

R Well I think, uh, the thing I like best, like I was saying before, is the higher pay, and, uh, uh, the security of getting a job in management

- 16 I Uh hum—you mentioned higher pay and the security of the job—are there any other kinds of things you think of there?

16 This type of probe was discussed earlier—the summary of conversation, then the request for other responses

R Well I guess you could say I like the supervisor's kind of work

- 17 I Uh hum

R It gives you a chance to work with the men and at the same time (interruption)

- 18 I You say you like the supervisory kind of work. Could you tell me a little bit more about what you have in mind there?

18 The interviewer felt that the respondent had not given enough information on the supervisory work. Hence the probe, which directed the respondent to this topic. However, the interviewer did not allow the respondent to exhaust

is, uh, handicapping you in your work Uh
what else do you think of in this respect?

26 This was a poorly timed probe The respondent was talking about his supervisory problems at a level which would give real insight into his foreman's role and its problems The interviewer chose to focus the respondent on a new area rather than to follow up on the basic problems

R Oh, I don't think there's anything else
I really ought to mention

27 I I'm interested in what you have in mind
there

27 Here again the respondent shows some resistance In this case the interviewer merely asks him to talk about these resistances In response, the respondent mentions a problem area This was probably a more effective technique than making a more supportive remark

R Well, they have awful tight production
schedules here

28 I Uh hum, and that affects you

28 This follows up the previous comment Here the interviewer recognizes the attitudes implied by the response It is a statement rather than a question

R The, uh general foreman holds our section responsible for getting out a certain amount of track each day It seems as if he doesn't know anything but just 100 percent all the time

29 I This causes you some problems, I gather

29 This again recognizes the attitude underlying the remark Notice that the response is in terms of the attitude rather than content The real attitude comes out in the response after

- 23 *I* Sure, I understand that, uh, what we were thinking of, uh, that uh, on most jobs a person has, there are some things that he may not like as well as others, some things that he may actually dislike. We are trying to get a general picture of, uh, what some of the things are that aren't so good.

23 This probe follows the resistance which the respondent showed in his previous remark. It is a general restructuring and is supportive in that it recognized the respondent's reluctance to be critical and attempts to make it acceptable for him to give negative statements.

R Well, one of the things you might say that's bad about this job is the condition we get the parts in.

- 24 *I* What, uh, what's that?

R Well, what I mean is, these steel cleats that we have to put on the tracks, we bolt them on, and there's another section further up the line where they're supposed to drill holes in the right places for us to slip the bolts in and half the time they do such a sloppy job there that when we try to put the bolts in place, we find they don't fit and we have to spend time reaming out the hole, and when we do that, we slow down on production and then the general foreman comes around and chews me out.

- 25 *I* Uh-hum.

R And, uh, it seems to me that if they got things tied together better, that wouldn't happen. We don't have a chance to talk that over.

- 26 *I* I see, uh, you've, uh, you've mentioned one thing, that the, uh, way you get the parts

change the question but merely defines what is meant

R Yeah, we have em

- 34 I Well, how, uh, *How do you get along with the shop steward?

R Oh, pretty good I guess I don't quite know what you mean 'get along'

- 35 I Oh, I suppose what we have in mind there is that we are interested in finding out how people perform these jobs, and how people in union shops who do these jobs get along, when they have to work together like this. What are your ideas on that?

35 It appears that the interviewer was caught off balance by the respondent's question. He was not sure whether the respondent was resisting or merely unclear as to what was wanted. He responded in terms of the question objectives.

R Well, most of the time, we don't have any trouble with each other. I try to take care of my job and he takes care of his. Of course, sometimes there are differences that have to be settled.

- 36 I How do you handle differences when they come out?

36 This is a direct question which is in line with the objective of the questionnaire item.

R Well, like suppose a man figures his job's timed too tight. He can mention it to me directly if he wants, or he can take it to the shop steward. Now, if the steward gets it, he can come around and we'll just talk it over, or if he wants to be nasty about it, he can file it as a formal grievance.

- 37 I I see. How do you usually work out these cases in your section?

Question 30 Those readers familiar with the principle of client centered therapy will recognize the technique This is the first place in the interview where there is real emotional content to the responses The recognition of that emotional content helps to bring the interview to a level of discussing those attitudes rather than conversation at the symptom level

R You're darned right it does—especially, like I was saying when the parts we get aren't quite right

30 I Yes, I see

R It seems to me that people higher up in management ought to find out more about how things are for us!

31 I You feel it would be an easier job for you if people higher up knew more about your job?

31 Again this brief summary helps the general permissive atmosphere

R Yeah, they don't come around to find out how things are really going—it's just 100 per cent production, or else!

32 I Yes, I see Well, let's turn to something else a minute—uh, *Do you have a shop steward in your section?

32 By asking this question at this time, the interviewer closes off a fruitful area of attitudes It is interesting to note that at this point the respondent shows negative reactions He fails to understand the interviewer's question and quibbles over words This may well represent his resentment at being closed off

R You mean, uh union?

33 I Yes, I mean a union shop steward

33 This clarifies the question It does not

general information about you, for example, your age—how old are you?

40 Having met resistance, the interviewer goes into more detail in describing reasons for collecting personal data. It is common in the research interview to get some resistance on these questions because of the concern over personal identification. Interviewers usually use the technique illustrated here. They make a short statement and ask the first question. If this brings resistance, they give a more complete statement of purpose.

R Well, I m 33

- 41 I *How many grades of school did you finish?

R Well, I never got a chance for much school

- 42 I Uh hum, about how far did you get?

R Eighth grade

- 43 I Eighth grade

R Had to go to work

- 44 I I see. Now that last one—*About what would your total income be this year for yourself and your immediate family?

R I don't see what that's got to do with it.

- 45 I Well, this is another one of these, uh items I was mentioning to you—people with different pay rates and salaries may very well feel differently on some or a lot of these questions we ask—for example, you remember our discussion about how you felt about your pay a while back. Well, it may be very well that a person getting one amount of pay would feel very differently from a person getting a different amount. This gives us a chance to make a statistical kind of analysis.

R Oh, he usually comes around and tells me what he thinks, and if we can work it out together, uh, we don't make a grievance out of it—grievances are just tough for the union and tough for us

- 38 I Uh hum, then in most cases you are able to work this out between yourselves

R Yes, he's reasonable. He's a little stuff necked sometimes on the time study things. Heck, I don't do the time study either. I'm in the same box he is

- 39 I Yes, I see. Well this covers about all the questions I wanted to ask you there. There are, there is just a little information I would like to get from each person we interview—uh, *About how old are you?

39 *At this point the interviewer is ready for the personal data information. He restructures before asking the personal data*

R I thought that, uh, these were gonna be anonymous and nobody was gonna care who gave the information

- 40 I Yes, that's right. Let me tell you a little bit about what we get here. As I mentioned earlier, before we started the interview, uh, we don't take people's names. We aren't interested in identifying them at all. We do, however, want to know something about the people we talk to because, you see, the older people who have been in the company longer may feel differently from the people who have been here a short time, and the younger people may feel differently from those who are somewhat older, things of that sort. We are not interested in identifying you in any way. And so, I have just a few questions of this sort. Would you just give me a little

ing seems apparent and simple—that is how to ask questions of people and get information from them. What is lacking is an understanding of the characteristics of a good interview—that is, the requirements that need to be met before an interview can be considered good. What are the principles of standardization, of validity etc. which we are trying to achieve?

Describing the goal for the interviewer can be done partly by helping him to understand the total research process and the part which he has in that process. The interviewer needs to know how a study is designed, the general principles of sampling and how the data are to be analyzed. This information will serve as a basis for his understanding of the interview in relation to the total research process. These research principles establish the basis for the interviewer's job. If this orientation is successful, the new interviewer now sees what he is trying to accomplish through his training. Furthermore, this knowledge provides him with a basic understanding so that he sees why he is trained in certain techniques.

The second training aim is to motivate the interviewer. We have implied in the discussion of 'goals' that the interviewer has some real reason for wanting to reach this goal. It is unwise, however, to assume that because the goal has been pointed out, he is highly motivated to achieve it. The interviewer must feel that what he is doing is important and significant—he must have an enthusiasm for his work. Although this is the usual part of a training function, it is well to point out that it is important to stress motivational aspects, for example, pointing out to the new interviewer why the study which he is about to undertake is important, what its function is how it will be used, why it is necessary that the data be collected accurately, and things of that sort. Another motivational factor which is common in interviewers is craftsmanship—that is satisfaction with an interview well done, particularly if the situation has been a difficult one. Earlier we talked about motivating the respondent to communicate. It is clear that it is difficult for the interviewer to motivate the respondent if he himself is not motivated.

The third aspect is training in interviewing skills, or imparting to the interviewer the specific methods and techniques which will make him an adequate interviewer. It is the feeling of the writers that in many interviewer training programs too much training is given in terms of 'rules' and specifics—that is in terms of 'The

45 *Again a statement of purpose was necessary. It is at this point that the new interviewer is likely to get on the defensive and have the respondent refuse to give an answer. A calm statement of purpose usually overcomes this. This respondent showed more resistance on these items than is usual. The reason may be that he gave considerable critical information during the interview. (Much of this does not show in this excerpt.) He would be concerned about the possibility of having his responses reported to the company and identified with him.*

R Well, I get seventy three dollars a week

46 I Seventy three dollars a week I see

SOME PRINCIPLES OF INTERVIEWER TRAINING

So far this chapter has stressed the point that the collection of data by means of personal interviews is a highly complicated technical job which demands much of the interviewer. It is clear that, if the techniques described here are to be effective, interviewers need careful training. Much of the validity of the obtained data depends upon the skill with which the techniques are applied, which in turn depends upon how well the interviewer is trained.

This section presents some general training principles which have been found to be effective. The training program has three major emphases. The first is to clarify to the trainee the goal of interviewing. In many training programs for other kinds of skills the goal of training is clear and the training need not, therefore, be greatly concerned with this aspect. If, for example, we are training a person to operate a typewriter efficiently, it is clear to the trainee that his task involves rapid and accurate manipulation of the typewriter. For the lathe operator, the same goal is evident. The goal for the interviewer is not so apparent. Most people have had some experience in interviewing, whether in the formal sense or not. In our everyday lives we often ask people questions to get information of one sort or another. To the new interviewer, then, the goal of train

the interviewer used techniques which were irritating or embarrassing. By analyzing his own reactions to being interviewed as he experiences the effects of the interviewer's techniques, he can become sensitized to the reactions of respondents. The trainees who are observing have a chance to see the performance and eliminate errors in their own interview techniques. Barron (4) makes this statement about the use of role playing as a device for training interviewers:

The use of role-playing or reality practice is being increasingly recognized as an effective means of translating the principles into methods, of learning the "how," of getting the feel of doing something in a situation where one is not playing for keeps. In training which is directed toward improving skill in interpersonal relations, it is offered as an effective way of bridging the gap between the formal study of principles, methods and techniques on the verbal level and actual work with these methods and techniques. It offers an opportunity for practice in the kind of work like interviewing, where close supervision and training on the job are very important.

In addition to the use of role playing as a device for transmitting skills, the use of phonograph recordings which illustrate various aspects of the interview and typical examples of interviews are very useful. They serve to point out to the interviewer what an actual interview sounds like and how an experienced interviewer handles a specific situation. They serve, too, as a basis for general discussion of interviewing methods.

No matter how effective the training, it is unrealistic to expect the original training to make finished interviewers or that it will be equally effective with all interviewers. One of the essentials of training is that further training be conducted periodically as the interviewers proceed in their work. As the interviewers grow more proficient, they become more interested and more involved in the fine details of the interviewing process. They want to discuss specific types of probes, the motivation of difficult respondents, etc. In such sessions, role playing is a valuable technique. It permits an interviewer who has a problem to act the role of the respondent and thus portray the difficulties which he is having. The difficulties can usually be ironed out through the role-playing session, which

first thing you should do is this The second thing you should do is that,' 'When you get a question of this sort, it should be handled this way, etc The interviewer is presented with a long list of specific techniques which he is to use, but the specifics fail to add up to a general, integrated system or a conceptualization of interviewing This conceptualization can usually be developed in discussions of the research process and the role that interviewing plays in the research process, by showing how the other phases of the research process are dependent on the interview, and by demonstrating how failure to follow these principles leads to error or invalid results Once this has been established, training on the specifics is in order Methods of skill training have been summarized by Bavelas (3) in the following statement What appears to be the most effective method of training skills is a common sense one—watch others, let us watch you, discuss and evaluate differences, and try again ' This implies the use of informal group discussion techniques and practice rather than lectures or dependence upon written material

One way of giving experience is to have the interviewer conduct actual interviews This, however, has the disadvantage that the trainer has only a second hand report of what took place during the interview, since he was not present at that time The ideal method is one in which an actual interview can be conducted in the presence of the trainees for all to observe and for all to discuss One technique which fulfills this objective and has other advantages is that of role playing or 'reality practice ' It has been adapted to serve as a technique for training in behavioral skills, primarily skills involving interpersonal relationships

In using role playing, one member of the group plays the part of the respondent, identifying himself with some person he knows and responding to the interviewer in terms of the role which he is playing Another person plays the role of the interviewer The rest of the group act as observers When the role playing session ends there is general discussion of the techniques that the interviewer used Many times the trainee gets as much out of playing the role of the respondent as he does out of playing the role of the interviewer By playing the role of the respondent, the "respondent" can perceive where the interviewer failed to get information and where

attention to rapport building 'probing' and recording of responses. These techniques are also illustrated in the sample interview and accompanying commentary.

BIBLIOGRAPHY

- 1 Adorno T W *et al* *The authoritarian personality* New York Harpers 1950
- 2 Bancroft G and Welch L H Recent experience with problems of labor force measurement *J Amer Stat Assoc* 1946 41 303-312
- 3 Bavelas A Role playing and management training *Sociatry* 1947 1 183-191
- 4 Barron M Role practice in interview training *Sociatry* 1947 1 198-208
- 5 Blankenship A B (ed) *How to conduct consumer and opinion research* New York Harpers 1946
- 6 Cantril H *Gauging public opinion* Princeton Princeton Univ Press 1944
- 7 Cartwright D P Some principles of mass persuasion *Hum Relat* 1949 2 253-268
- 8 Crutchfield R S and Gordon D A Variations in respondents' interpretations of an opinion poll question *Int J Opin and Attitude Res* 1947 1 No 3 1-12
- 9 Festinger, L Back K Schachter S Kelley H H and Thibaut J *Theory and experiment in social communication* Ann Arbor Edwards 1952
- 10 Hyman H H Problems in the collection of opinion research data *Amer J Sociol* 1950 55 362-370
- 11 Jahoda M Deutsch M and Cook S W *Research methods in social relations Part I Basic processes* New York Dryden Press 1951 Chap 6
- 12 Kinsey A C Pomeroy W B and Martin C L *Sexual behavior in the human male* Philadelphia Saunders 1918 pp 35-63
- 13 Lazarsfeld P F The controversy over detailed interviews—an offer for negotiation *Publ Opin Quart*, 1944 8 38-60

also provides the opportunity for the rest of the group to profit from his experience and to learn along with him

SUMMARY

The purpose of the present chapter has been to discuss the technique of the research interview, to offer a rationale or theoretical framework for the technique described, and to put the interview in perspective as one of the various devices for data collection which are at the disposal of science

We began with the postulate that scientific progress depends importantly upon the systematic collection of data and that this involves (1) a statement of specific research objectives, (2) definition of the data required to meet such objectives, (3) determination of the population from which these data can be obtained, (4) selection or development of techniques adequate to evoke the data. We have attempted to demonstrate that the interview can approximate these criteria in social research, and that the interview is especially adapted to the collection of data about attitudes and perceptions, beliefs, feelings, past experiences, and future intentions

The problem of respondent motivation was discussed in terms of two major motivational sources: (1) the respondent's perception that by participating in the interview he may help to achieve some goal or bring about some change which he considers desirable, and (2) the direct gratification or catharsis which the respondent realizes by speaking to a person who is understanding and accepting of his opinions and ideas

Questionnaire design was presented as the task of creating an instrument which would serve to translate the research objectives without bias into terms understandable to the respondent and would, at the same time, assist rather than retard the interviewer in motivating the respondent to communicate. The specific aspects of questionnaire construction were also presented, including language, frame of reference, information level, social acceptance, wording, and question sequence

The specific techniques which the interviewer must employ to evoke complete and frank responses were reviewed, with especial

attention to rapport-building, "probing," and recording of responses. These techniques are also illustrated in the sample interview and accompanying commentary

BIBLIOGRAPHY

- 1 Adorno, T W , et al *The authoritarian personality* New York Harpers, 1950
- 2 Bancroft, G , and Welch, E H Recent experience with problems of labor force measurement *J Amer Stat Assoc* , 1946 41, 303 312
- 3 Bavelas, A Role playing and management training *Sociatry*, 1947, 1, 183 191
- 4 Barron, M Role practice in interview training *Sociatry*, 1947 1, 198 208
- 5 Blankenship, A B (ed) *How to conduct consumer and opinion research* New York Harpers 1946
- 6 Cantril H *Gauging public opinion* Princeton Princeton Univ Press 1944
- 7 Cartwright, D P Some principles of mass persuasion *Hum Relat* , 1949, 2, 253 268
- 8 Crutchfield, R S , and Gordon D A Variations in respondents interpretations of an opinion poll question *Int J Opin and Attitude Res* , 1947 1, No 3, 1 12
- 9 Festinger, L , Back, K , Schichter, S Kelley H H and Thibaut J *Theory and experiment in social communication* Ann Arbor Edwards, 1952
- 10 Hyman, H H Problems in the collection of opinion research data *Amer J Sociol* , 1950, 55, 362 370
- 11 Jahoda, M , Deutsch M , and Cook S W *Research methods in social relations Part I Basic processes* New York Dryden Press 1951, Chap 6
- 12 Kinsey, A C , Pomeroy W B , and Martin C L *Sexual behavior in the human male* Philadelphia Saunders 1918 pp 35 63
- 13 Lazarsfeld, P F The controversy over detailed interviews—an offer of negotiation *Publ Opin Quart* , 1944, 8, 48 60

- 14 Lazarsfeld P F *The art of asking why* *Naa Marketing Res*, 1935
1 26 35
- 15 Parten M B *Surveys polls and samples* New York Harpers 1950
- 16 Payne S L *The art of asking questions* Princeton Princeton Univ
Press 1951
- 17 Rogers C R *Client centered therapy its current practice, implica
tions and theory* New York Houghton Mifflin 1951
- 18 ——— *Counseling and psychotherapy* New York Houghton Mifflin
1942

Observation of Group Behavior

Roger W. Heyns and Alvin F. Zander

Within the past few years there has been a great increase in the use of observation methods in the study of social phenomena. These experiences have indicated that direct observation of social behavior can provide reliable and conceptually meaningful data in field studies as well as in laboratory experimentation.

This increase in the use of observers has been accompanied by an increase in methodological sophistication in observation methods. There is a growing awareness, for example, of certain typical problems in the development of observation schedules, the training of observers, and the achievement of reliability. Many of these problems have not yet been subjected to methodological research, but there is a good deal of "wisdom" in these areas which, until the necessary research is done, will help the investigator to avoid some of the more common pitfalls.

This chapter will deal with two principal types of observation instruments: category systems and rating scales. We shall discuss the finished products of both types and the problems involved in their development. To provide a focus for this discussion, we shall begin by describing an observer team in an actual situation.

AN OBSERVER TEAM IN ACTION

The Setting

Let us suppose that an observer team of two is studying methods of problem solving in groups. Let us suppose, further, that this phase of the investigation involves the observation of a large number of groups in the field. This means that the observations are being done in a relatively uncontrolled situation. The team can control neither the kinds of variables which will be present nor the behavioral responses to them. The specific duties which have been assigned to the members of this team as well as the way in which they are to perform them have been dictated by the study design and the theoretical framework which have been developed.

The Entrance and Behavior of the Observers

The observers have been trained to make themselves as non-threatening as possible. Thus they have made sure that the group knows ahead of time that they will be present and that the group is generally aware of the observers' purposes in attending the meeting. They arrive at the meeting place early, introduce themselves to the chairman, and answer any additional questions he may have concerning their purpose and the way in which they will function. Their manner toward both the chairman and the group is positive, understanding, and supportive. Both their style and the content of their replies to any questions are intended to make the group members feel that the observers are present to record objective facts rather than to evaluate the quality of their meeting or the performances of any individuals.

Before the meeting begins, one member of the observer team explains the general purpose of their research, stressing the fact that they are interested not in the content of the discussion but rather in the ways in which groups go about solving their problems. He suggests that the members of the group disregard the observers and not involve them in the discussion. He also describes briefly the nature of work they will be doing while the meeting is going on.

The observers take especial precaution to keep their activities from interfering with the meeting. They sit as far away from the

group as space permits. They avoid conversation with each other and any other communication which might reveal their attitude toward what is occurring in the meeting unless there is laughter, applause, or some other reaction which they can appropriately make without appearing to differ from the dominant climate in the room at the time.

The Data Obtained by This Team

Observer *A* codes the problem solving process of this group; observer *B* records the content of the meeting. To test the hypotheses developed for this study, the problem solving observer uses a prepared category system.¹ He is responsible for coding each relevant contribution of each member into one of a number of categories. He also records which member made the contribution and to whom it was addressed. These categories are listed on a standardized form which facilitates rapid recording of these data. This observer watches the group interaction in terms of the categories which are listed below with their definitions. By the time the observer is on the job, he has memorized these definitions, of course, and hence they are not repeated on his observation form.

PROBLEM SOLVING CATEGORIES²

Goal setting. These contributions have the function of establishing or suggesting goals or objectives both procedural and content. They are concerned with ends to be attained. These objectives, goals, or ends may be those of the individual which he is trying to have the group attain; they may consist of statements of accepted goals of the group or part of the group.

Problem proposing. These contributions serve the function of presenting a problem either in content or in procedure. They are concerned with means to ends or goals.

Information seeking. These contributions have the function of seeking to obtain information of an objective, factual, or technical nature. The information sought is from the area of

¹ At some point in the data collection process an additional observer can be introduced to provide a reliability check for *A*'s coding.

² This set of categories and other examples in this illustration are taken from the procedures used by the Conference Research Project, University of Michigan (8).

fact on which the group decision is to be based or which has bearing on the decision. Contributions seeking factual, objective, or technical information concerning the procedure of the group or an individual are classified here.

Information giving These contributions have the function of providing objective, factual, or technical information either in the subject area or with respect to procedure. The category includes the citing of examples or illustrations.

Solution proposing These contributions serve the function of indicating solutions to problems. They are suggested means to ends. Modifications of or additions to solution proposals previously offered are classified in this category if the context gives the contribution a solution proposing function.

Development seeking These contributions serve the function of attempting to obtain clarification of previous contributions. They seek to determine what was intended by a previous contribution, what its implications are, what inferences are permissible. These frequently take the form of an inference stated as a question.

Also included here are contributions which facilitate the procedure of the group by asking the group as a whole or individuals to comment, indications to individuals that they have the floor, etc.

Development giving Contributions here elaborate, make explicit, enlarge on contributions. Included here are inferences from previous contributions, self-repetitions or restatements by others of previous contributions, reflecting types of contributions which are distillations of previous contributions without intent to get clarification but which are, rather, declarative statements of what the previous contribution stated or implied. Finally, this category includes contributions which provide the rationale, reasons, or arguments for the individual's positions. They give his reason for his saying what he does.

Opposing These contributions are characterized by an opposition to, resistance to, or disagreement with a suggestion, solution, interpretation, etc. Responses which point out obstacles, difficulties, or objections are included here.

Supporting These contributions serve the function of indicating agreement or approval of a suggestion or solution proposal. Included here are indications of approval of the fact that

another has contributed, whether approval of content is present or not. This is a supporting comment in procedure.

Summary seeking These contributions ask, in effect, for a summary—e.g., 'I'm lost, where are we now?'

Summary giving These contributions summarize the group's progress to date. They refer either to substantive material discussed over a period of time to conclusions reached, or to the group's procedure over a period of time. Summary statements of individual participations are not included here.

Non problem directing This category includes irrelevancies of the tangential sort and a myriad of responses of an interpersonal sort, such as 'Give me the ash tray' and 'How about opening a window?'. It includes statements which have no reference to the subject matter of the conference or to the group procedure.

While observer *A* notes problem solving contributions in the categories above, observer *B* records the nature of the meeting content. He keeps a running account of the actual subject matter discussed in accordance with the following:

- 1 Notation of each topic discussed
- 2 Classification of each topic into one of the following
 - a Procedural a topic having to do with the procedures or process of the group
 - b Substantive a topic having to do with the subject matter of the meeting
- 3 Notation of the nature of the task confronting the group at each point on its agenda, using the following categories
 - a To arrive at a decision
 - b To approve a decision already made
 - c To receive or give information
- 4 Observation of what actually happens to each item on the agenda by noting whether the agenda item was
 - a Completed
 - b Postponed
 - c Left uncompleted

For each of the tasks above the observer has been trained to follow carefully stated definitions in an 'Observer's Manual'.

When the meeting is finished, each observer is required to rate certain aspects of the meeting particularly the areas of communication, motivation, and interpersonal relations. Each works independently, using the following scales

- 1 *Understandability* To what extent were the participants getting the meaning of one another's statements?

0	1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	---	----

They were talking past one another, there was much misunderstanding

Communicated directly with one another

(Note This rating should include not only trouble with particular words but also more general conceptual processes such as difficulties in level of concreteness style of expression etc)

- 2 *Opportunity to Communicate* To what extent did the participants have opportunity to talk?

0	1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	---	----

Never had opportunity to talk

Seldom had opportunity to talk

Usually had opportunity to talk

Had every opportunity to talk

(Note In some groups this can be judged by the number of times persons seemed eager to get the floor but could not and the number of simultaneous participations. In other groups the participants have already learned not to try to talk because of the dominance of a few members hence, although things need saying the participants have little opportunity to say them)

- 3 *Ego involvement* How much did the members have at stake in the problem-outcomes?

0	1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	---	----

Nothing to gain or lose

Much to gain or lose

- 4 *Urgency* How urgent did the group regard these problems?

0 1 2 3 4 5 6 7 8 9 10

No urgency

Very pressing

(Note This is a judgment of the extent to which the group felt it was necessary to arrive at a decision at this meeting)

- 5 *Importance* How important to their organization were these problems regarded by the group?

0 1 2 3 4 5 6 7 8 9 10

Of little
consequence

The very life of
the organization
depends upon it

- 6 How formal were the interrelationships among the people in the group?

0 1 2 3 4 5 6 7 8 9 10

Extremely
formal

Predominantly
formal

Largely
informal

Completely
informal

(Note Rate on basis of mode of address number of personal comments and number of asides used by the individuals indicating social distance among them)

- 7 How supportive and accepting was the group of its members

0 1 2 3 4 5 6 7 8 9 10

The group was
highly critical
and punishing

The group was
permissive and
highly receptive

- 8 How pleasant was the affective interpersonal atmosphere of the group?

0 1 2 3 4 5 6 7 8 9 10

Very unpleasant,
quarrelsome, critical
and unfriendly

Very pleasant,
personable, warm,
and enjoyable

Below some of these scales there are instructions to raters concerning the cues which are important and the extent to which certain factors should be weighted in arriving at a final rating. More detailed instructions of this sort for each item are again provided in an "Observer's Manual."

Each observer also writes anecdotes describing incidents or conditions which he thinks might be important. Specifically, he is asked to note factors which have a bearing on the observations obtained (their validity, interpretation), factors which should be included in subsequent observations, and the presence of conditions which might make the group unsuitable for inclusion in the final analysis of data.

This brief description of the steps a typical observer team might follow has been placed in a field setting. The observation process in a laboratory is not very different, however. The experimenter would probably introduce the observers to the persons being observed and describe the activities of the observers. More will be said about the introduction of observers and the problems of observer decorum in a later section of this chapter.

Now that we have described the activities of a hypothetical observer team, we are ready to turn to a discussion of the technical features involved in the nature and development of such a system.

CHARACTERISTICS OF AN OBSERVATIONAL SYSTEM

This section will be concerned, first, with the definition and description of two major types of observation systems: category systems and rating scales. The descriptions present the characteristics of finished products of both types. Later we shall discuss problems in the development of these procedures.

Category Systems

One of the most useful devices to describe qualitative social situations in quantitative form is that of coding the behavior within separate categories. For the purposes of this section, a category is a statement describing a given class of phenomena into which observed

behavior may be coded, a category system consists of two or more categories. A carefully developed category system provides a common frame of reference for observers and increases the likelihood that the relevant aspects of the total behavior will be noted with reliability.

The exact nature of the category system—i.e., its characteristics with respect to the number of categories, the level of conceptualization they involve, the applicability of the set to a wide variety of situations, etc.—depends upon the purposes of the investigator and the theoretical framework within which he is working. Although it is possible to distinguish various types of category systems along a large number of dimensions, several dimensions seem to us especially useful for understanding the kinds which have been used and the types of data which can be obtained by their use.

THE DIMENSION OF EXHAUSTIVENESS Some category systems are developed in such a way that all the behavior observable can be classified into one of the categories in the set. Bales' interaction categories make up such a system (2). The 12 categories in his system have been developed with the objective that all verbal behavior in a small face to face group be codable into one or another of them. A contrasting system is that used by Jack in her study of ascendance and submission in a play situation (9). Her set of categories focused on forms of ascendant behavior, behavior not in these categories was not coded. In a sense, of course, all less than exhaustive category systems are really exhaustive, since behavior not codable into one of the categories is implicitly in a category labeled "not in the system." This distinction is a real one and consists essentially of determining how much of the total observable behavior is to be classified into a defined category.

The question of whether or not a category system ought to be exhaustive (in this sense) must be decided by the experimenter in the light of his purposes. Two considerations seem worth pointing out, however. The first is that much time can be saved in analysis if the instrument contains only those categories which are necessary. The second consideration concerns the nature of the behavior not categorized when a less than exhaustive category system is used. Even though no further discrimination is necessary than "it is not in the system," it may nevertheless be important to know the total amount of behavior in this residual, undifferentiated category. This can be

done only if one has a record of the amount of total behavior which was not specifically categorized

THE DIMENSION OF INFERENCE Another way of differentiating category systems involves the amount of inference that they require of observers. Some category systems require that the observers coding the behavior proceed from the actual behavior which they noted to a deduction about this behavior. At the other extreme are procedures which require the observers to place the behavior they saw or heard into categories with no requirement of inference. Let us suppose that an experimenter is specifically interested in the effects of strict disciplinary practices in the classroom on the kinds of interaction of children on the playground. At the simpler extreme, he might develop categories such as 'Shoves other children,' 'Calls other children names,' 'Asks for help.' These behaviors are selected either because they represent the level of data the experimenter needs to test his hypothesis or because they permit later reclassification into categories which are at the appropriate level. On the other hand, he may use such categories as 'Shows hostility,' 'Demands submission,' 'Desires support,' 'Resents dependence.' In the latter case the observers are looking at the same behavior but are making inferences concerning it. The material which is coded is based upon these inferences.

When some inference is required to test the hypotheses, the essential difference between the two category systems is who makes the inference. In either case, some theoretical system is required on the basis of which inferences are made, either by the experimenter when he confronts the data presented to him by observers or by the observers when they observe the behavior. In other words, in the first case, the observers note the incidence of behavior in categories and the investigator makes the inferences during the analysis of the data, in the second case, the experimenter asks the observer to make inferences after he has instructed him as to the kinds of inferences that can be made. This instruction actually consists in describing the behaviors which can be placed in each category.

Once again the decision as to how much inference to require from observers depends in large part on the purposes of the experimenter. There are, however, a number of considerations which deserve special mention. One of these is the degree of confidence one has in the clarity of the concepts being used and the behavior

which may be described in terms of these concepts. For example, when we ask observers to make inferences concerning the motives or emotional state of the actor from behavioral acts, we must either have empirical evidence that a number of phenotypically different behaviors have the same genotypical dimension or that our theory about this situation is such that they may have. The availability of sophisticated observers is another factor which affects the advisability of using category systems which involve high level inference. If one's resources in this respect are limited, it seems much more desirable to define clearly the categories of behavior to be classified with a minimum of inference, leaving the inference task to the experimenter.

NUMBER OF DIMENSIONS Some category systems are developed within a single frame of reference, they require the observers to focus on behaviors which are, at one level of conceptualization homogeneous. Others include categories which describe social behavior along a number of dimensions. This distinction can also be applied to single categories within a system. An example of a category system focusing on a single aspect of group process is that developed by Heyns and Berkowitz (8) for the description of the problem solving process in decision making groups (described at the beginning of this chapter). Each participant's contribution is classified into one of 12 categories on the basis of the problem solving function performed by that contribution. Other dimensions of group interaction, such as the emotional impact of contributions are ignored or only minimally represented.

Bales' (2) interaction category system is an illustration of a category system involving more than one dimension. For example Category 1, 'Shows solidarity,' and Category 2, 'Shows tension release,' seem to be descriptions of interaction along affective dimensions. Categories 5, 'Gives opinion' 6 'Gives orientation,' 4 'Gives suggestion' refer to the intellectual problem solving activity of the group.

There is, again, no pat answer to the number of dimensions which should be attempted in a category system. It should be pointed out, however, that the number of dimensions dictates the number of frames of reference which the observers must be aware of and must use. A large number tends to reduce agreement between observers. A further complication in the use of systems involving more

than one dimension is that the categories on the different dimensions may not be exclusive. Thus, for example, at one level of analysis a question may be coded quite correctly as a request for information. At the very same time, it may be coded equally correctly as an expression of hostility. This difficulty can be overcome, of course, by permitting multiple coding. This may not be desirable for other reasons, however. Finally, when a single category includes behavior on more than one dimension, there should be strong theoretical or empirical support for the proposition that at another level of conceptualization these two dimensions are similar. Without this, a category score would have no single meaning.

Discrete vs. Continuous Categories

Some category systems are so constructed that the categories within them can be placed along a continuum. Others, even though they have only one dimension, contain discrete categories, which cannot be located in relation to each other on any very clear continuum. We can illustrate the first type by referring again to the researcher interested in describing the playground behavior of children. Let us suppose that he has a number of categories into which aggressive behavior can be coded. Category 1, 'Mild verbal attack,' Category 2, 'Verbal threats and threatening gestures,' Category 3, 'Direct physical attack.' According to one classification scheme, this would be a continuous category system, ranging from mild aggressive behavior in Category 1 to severe in Category 3. The adequacy of a continuous system rests on the adequacy of the theoretical framework. In the example given, a threat is assumed to be a more severe form of aggression than verbal attack. There might be situations, however, in which this was not true. When assumptions concerning the ordering of categories are justified in a given situation, continuous category systems are very desirable. When developed adequately, they constitute a scale (see Chap. 11).

Discrete category systems contain categories which have no such relationship to one another. An example of this type is the problem solving category system referred to earlier. It is not possible to locate such categories as supporting, solution proposing, goal setting, and developing on a single, clear continuum.

Rating Scales

Simple rating scales are also often used to record quantified observations of a social situation. They may be used to describe the behavior of individuals, the activities of an entire group, the changes in the situation surrounding them, or many other types of data. Rating scales often provide more superficial and less reliable data than do well-developed category systems such as those just described. However, practical limitations may force one to rely upon this method to guide observations.

By the phrase "simple scale" we mean a scale with a set of points which describe varying degrees of the dimension being observed. Observation rating scales have seldom been submitted to rigorous "scaling" treatment in their development, probably because it is often difficult to get a sufficiently large number of trials with any one observation schedule.

HOW RATING SCALES ARE USED. Rating scales are most often used in either of two ways: (1) to record behavior at frequent intervals throughout a sample of social interaction, or (2) to rate the nature of the entire social event after it has ended. An example of the former type was used by Lippitt and Zander (10) in a field experiment on Scoutmaster training. Observers noted the behavior of Scoutmasters in Scout meetings before and after they had received training. One five-point rating scale was used to rate the physical symptoms of group tension shown by the boys during meetings led by the trainees. The scale was marked whenever the program activity or the group atmosphere changed during the meeting. The description of the scale and the observer instructions was as follows

Physical symptoms of group tension This scale is concerned with the state of tension existing in the group as it is revealed through physical symptoms. It will indirectly measure the amount of psychological strain created in the boys during various parts of a Scout meeting. The assumption is made that when a boy is physically tense to a noticeable extent there is a corollary psychological tension. This psychological tension may be rebellion or fear, or it may be tenseness caused by the desire to reach goals set before the group (by the Scoutmaster, other leaders, or the boys themselves). It will be difficult sometimes to know if the

tension is goal directed anxiety or whether it is an emotional reaction. An anecdotal note on the source of tension will be valuable if discernible. This scale asks for both sorts of tension on the hunch that most of the ratings will concern goal directed tension and those which do not will be apparent from the rest of the data. According to physical signs, how tense are the boys at this time in the meeting?

0 position—*Can't rate*

1 position—*Very relaxed* The group is physically and psychologically taking it easy. This does not mean that they are in a state of rest or repose. It means that they are carrying on the kinds of activity which occur at a Scout meeting without any apparent air of tension. This may be a sprawled conversation group or it may be a relaxed game. The boys seem comfortable, they act and look the way people do after they rise from a good meal.

2 position—*Relaxed* Mark this category if the group is relaxed but not as greatly relaxed as in the category above.

3 position—*Middleground* This category is marked if the boys are acting as most people do most of the time. The goal has a more positive valence than might be true in positions 1 and 2. There is a small amount of tension but it is not great enough to be expressed in physical signs of tension. They may be physically active or sitting still. Facial expression shows no apparent signs of strain.

4 position—*Restless* This point is marked to describe group behaviors which indicate psychological tension. Boys may be trying hard in a signaling contest or a written examination. Sometimes tension may be apparent in the purposeless movements by the boy (purposeless in the sense that they seem to have no relation to the group goal at the time). These are such movements as hand wringing, foot twisting, tongue chewing, drumming, and other nervous mannerisms.

5 position—*Keyed up* Here tension is very obvious. Its physical signs are clenched fists or hunched bodies or extreme

signs of restlessness or a tense anxious expression. A football crowd watching the kick for a point after touchdown would be keyed up. The keyed up behavior may be shown by boys who are in any posture or state of movement. During a fast basketball game for example the signs by which it is recognized might depend on the pitch and frequency of shouts, the facial gestures, etc. Boys who are required to sit through a tongue lashing may be keyed up but their tension might be shown in facial gestures and inhibited movements.

Let us ignore for the moment the adequacy of the assumptions on which this scale was based and the problem of reliability in observation on a rough scale such as this. It is clear that this is an attempt to get a measure of the significance of a large variety of bodily movements which are interpreted by the observers as indicative of psychological tension and that these many movements scattered throughout the group are gathered in a relatively economical way. An example of a scale used to record behavior at the end of a meeting was described earlier in this chapter. Another such scale was developed by Fouriez, Hutt, and Guetzkow (6) and was used to rate the amount of self-oriented need behavior shown by each member of the group. This type of behavior is described by the authors as follows:

[Self-oriented need behavior] is not necessarily directed toward a group goal or the satisfactory solution of the group's problems. It is directed primarily to the satisfaction of the need itself, *regardless* of the effect on the attainment of the group goal.

The scale has 11 points ranging from 0 (no expression of self-oriented need) to 10 (all behavior of the self-oriented need type) as follows:

- _____ 0 No expression of self-oriented need
- _____ 1 Some slight indication of self-oriented need behavior
- _____ 2
- _____ 3 Some self-oriented need behavior indicated but not prominent

preferable to record the frequency of each cue as it occurred. This could be a tremendous job, however, since there could be a large number of behavioral cues which might indicate the presence of one of the factors listed above. If one were interested in a large number of factors, the cues could mount to impossible numbers. Thus, rating scales, which contain a variety of behaviors at each point on the scale, are more efficient since they can provide more data per observer and more dimensions per unit of time. The observer using a rating scale is in the role of a human collating machine. He observes a number of acts throughout the group, integrates them in his mind, and makes a judgment as to which point on a number of scales best describes his interpretation of the varied behavior.

Thus, at some points in a meeting when events may be occurring so rapidly that it is impossible to record the nature of the activities of each person, it may be easy to record a summary statement along required dimensions by means of a rating scale. Indeed in some cases one needs measures of change in behavior only at the point where changes in other conditions occur. For example, the rating scale on physical symptoms of tension earlier described was used because it provided a quick snapshot of the state of affairs when the group changed program activities (perhaps from a teaching and learning condition to one in which the group was playing a game). Many new and important cues may happen rapidly under such a changed state of affairs (perhaps a sudden release of tension) so that one could not record them all. Nor can one depend on a leisurely tabulation of whatever it is possible to record within an observer's physical limitations since the program may rapidly change again before an adequate population of behavior cues can be tabulated. In such a case, ratings provide useful summary data.

Rating scales can be useful in the exploratory or pilot stages of a study. If one is uncertain what the cues are for a given type of behavior, scales can be useful in defining them. Let us say that one is interested in determining the nature of the signs which could be used in identifying supporting behavior. One might expect that the content of the words said by a person, the tones of his voice, the bodily postures, and his facial expressions might all reveal clues as to the presence or absence of supportive behavior. In this case, frequent ratings, with notations of the cues one used to make this decision, can be made by several observers. When they compare

their ratings, and the cues they used, they begin to be aware of the behavior they use as indicative of supporting behavior. Further refinement of the scale, another period of joint observations, and comparative discussion make increasingly clear the clues that are most indicative of this behavior.

Under ideal circumstances rating scales should provide data which are strictly comparable to those obtained by the use of a category system. This ideal is difficult to reach, however, since the variety of behaviors which one must include within any one rating may create problems of reliability. Then, too, a rating scale which is discriminating enough for one type of behavior may be too gross for another. If the scale is so gross that most of the data fall at one end of the scale, the user runs into problems in the statistical treatment and interpretation of the findings. Such problems as these are treated in more detail in Chapters 6 and 11, on the construction of measuring devices.

What criteria can one use to decide whether one should use single categories or rating scales? In general, whether one uses frequency counts or ratings depends upon one's resources and the demands of the problem. The greater the precision required, the less one is likely to use rating scales.

PROCEDURES IN DEVELOPING A SYSTEM OF OBSERVATION

Systematic observation of social behavior has a relatively short history. Our information is still too limited to give the researcher many rules of thumb which he can follow in the construction of his observational tools. We can, however, point out the major decisions which the experimenter must make, discuss some of the considerations which must be taken into account, and suggest some criteria which might be used in evaluating various alternatives. In fact, it is clear to those who have had considerable experience in this area that there is no single solution to each of the problems; the best solution is in terms of the objectives of the study.

The importance of theory in the design of experiments is often neglected. We need only point out here that the theoretical framework plays a central role in determining the decisions which the

experimenter must make in developing his methodology. This is particularly true with respect to selection of the behavior to be observed or rated and the definition of the categories into which these behaviors are to be placed. Without a knowledge of the purpose of the experiment or research and the theoretical setting in which the experiment takes place, no one can prescribe for an experimenter the dimensions with which he ought to be concerned, or the amount of inference he ought to require of his observers. The experimental design and the specific hypotheses to be tested will dictate the level of reliability of observed data which will be necessary. Many other decision implications of the theoretical framework could be stated, several of these will be made explicit in the discussion of decision areas which follows.

Some of the decision areas to be discussed are more pressing for rating scales than for category systems or vice versa. The problems are, however, sufficiently common to both techniques to warrant their discussion in connection with both.

The Frame of Reference

The frame of reference is involved, in part, in the problem raised in connection with the level of inference required from observers, and, in part, in the problems discussed under number of dimensions. We can explicitly restate the problem in this setting by saying that, in the interests of reliability, observers must be clear about the dimension to be observed and about the vantage point they are to use in observing, recording, coding, or rating it. Observers may be instructed, for example, to observe the social interaction along the dimension of interpersonal affect—i.e., the extent to which the members of the group are personally fond of each other. As the dimension is fairly clear, the question arises whether the observers are to react to the observed behavior as if they are participants or whether they are to identify with the actor, making inferences about his motives. These are different frames of reference. Many others are possible. Three of the more common frame of reference problems will be discussed.

THE CONTEXT AS A FRAME OF REFERENCE One of the recurrent sources of inter-disagreement and one which is often theoretically important is the extent to which the social context of the act or

behavioral unit is to be taken into account in the coding or rating. Few would seriously argue that the context (the situation in which the behavior being coded or rated takes place) should be ignored. It is rather a problem of what part of the context, how much of it, is to be taken into account. In the coding of the problem solving functions, for example, a specific contribution might serve to enlarge on or to develop a previous contribution. This earlier contribution might have been opposing a still earlier contribution. One coder might classify the present contribution as *developing*, whereas a second coder, having the larger context in mind, would classify it as *opposing*. The experience of some workers in this area indicates that, if theoretical considerations do not dictate the answer, using the most immediate situational context as the frame of reference makes for the greatest amount of agreement. This is especially true if interaction is rapid.

THE INTENT AS A FRAME OF REFERENCE Another prominent source of interobserver disagreement is the extent to which observers permit their judgment of the intent of the actor to color their ratings or categorization of his behavior. This is often a problem when it is the explicit purpose of the experimenter to ignore the objectives implied in a given statement. An illustration of this difficulty is that of the group member who is asking questions concerning the implications of a proposed solution to a problem. The observer oriented to ignore intent would classify this behavior as information seeking. The observer oriented toward taking intent into account might code this same questioning behavior as attacking or opposing, basing his judgment on the manner in which the questions were asked.

The experimenter may, however, be interested precisely in the intent of an act. When this is the case, two sources of observer unreliability may occur: (1) observers may disagree concerning the nature of the cues to use in identifying the intent of the actor, or (2) observers may disagree in the degree to which they take intent into consideration. In short, explicit instructions should be given observers concerning the cues to use for identifying intent and the extent to which judgments concerning intent or motivation for an act should enter into coding or rating.

THE EFFECT AS A FRAME OF REFERENCE A specific aspect of the context problem mentioned above is the influence of the effect of the

act on the coding or rating. A statement in a group discussion might be coded by one observer, ignoring effect, as a request for information. It may happen, however, that the following remark does not give the requested information. It is a defensive remark on the basis of which another observer judges that the original question was a critical or opposing act. Disagreement between raters or coders frequently results from a differential tendency to permit effects of acts, or perceived effects of acts, to influence their judgments. Whether or not this is to be permitted depends on the design and the interests of the experimenter. It is clear that this is an important decision area, certainly from a reliability standpoint and probably from a theoretical standpoint as well.

Size of Unit

After the experimenter has determined the dimensions along which he wants ratings or categorization, he faces the problem of determining the size of the unit of behavior to be rated or categorized. Failure to state clearly to the observers the precise limits of the unit of observation is a frequent source of unreliability. The reliability of the coding depends in part on the reliability of deciding what constitutes a unit of behavior.

Roughly speaking, the size of the unit may vary from a single act—for example, the contribution of a single participant—to a total period of interaction. The observers in the field studies of the Conference Research Project (8) made ratings of the adequacy of the problem solving process and used the entire conference period as the unit for some ratings, and single acts for others. Time sampling (categorizing or rating units of time distributed randomly over the total action period) is an intermediate technique. This can obviously include a single act or all acts.

In categorization, the selection of the size of unit is not independent of the category system itself. This has led to a fairly common practice of defining units in some such fashion as this: "any act which is classifiable into a single category." Failure to consider this aspect in the definition of units will perforce affect reliability. Thus, in practice, experimenters dealing with verbal behavior of members of problem solving groups have often found it impossible to deal

with a total simple participation as the unit. A lengthy verbal statement too frequently contains elements which permit classification in several categories.

Sampling Methods in Observation

It often happens that one cannot, or does not want to, record all of the behavior that takes place in a given situation. This may occur because the meeting lasts too long to observe continuously, too many persons are present, interactions come too rapidly to record, or for any of a variety of reasons. In such a case we are forced to develop a method of obtaining a representative sample of the behavior being observed. This can be done in a number of ways: (1) Attention can be concentrated on the behavior of a few of the members, ignoring the others present. (2) Attention can be directed to each person, or to a number of persons in the group, each for a given length of time. (3) The whole observation instrument can be divided into parts and the social setting can be observed in terms of each part of the observation schedule for a standardized length of time. (4) Observations can be made only when certain key behaviors have been introduced into the meeting. (5) The most frequently used system for obtaining representative samples of the behavior being observed is the time sampling system, in which a standardized time unit is selected during which observation takes place. The assumption is that these parts will be an adequate description of all of the events.

In general, it is risky to use sampling procedures unless one has an adequate theory to guide the selection of the sample. As an illustration, let us assume that we wish to observe a number of behaviors in a group but for some reason it is necessary to use a time sampling procedure. We make the decision to observe in terms of a certain set of categories for five minutes, skip ten minutes (during which we may be observing with a different set of "spectacles") and return for another five minutes with the original categories and so on. Any social situation, however, is a changing set of activities, and we may discover that certain events occur during the ten minutes we are not observing which distort the representativeness of the records made during our time samples. Thus, we would rightly be suspicious of the adequacy of our data. We could have

avoided this difficulty if we had had some notion of the range of possibilities of crucial changes in this group which could affect the adequacy of our time samples and had arranged our sampling procedure in such a way that it covered all of the crucial differences which may have distorted the representativeness of our time sample.

Only an adequate theory can help us anticipate what these changes are likely to be and whether they are likely to make any difference to the coverage of the sample. If we were interested for example, in the frequency with which group members on whom we have certain personality measures perform in a meeting according to certain predicted patterns, we might wonder whether it makes any difference if parts of the meeting are not included in our sample. If our theorizing leads us to suspect that certain types of group procedure change the predicted relation between personality measures and performance in a group, we must then be careful about time sampling. It could happen that the standardized units of time we choose to observe may not be a representative sample of types of procedures occurring in the group. Thus the data would be biased by the particular time units studied. It is safer, then, to use time sampling only when we have a theory to guide the observations, since the theory can help us to decide whether important parts of the data are being buried or are being recorded in the most effective fashion.

Sometimes it is possible to determine in a pilot study whether time samples of behavior are representative. The staff of Conference Research (8) found that the distribution of problem solving behavior from one five minute period to the next fluctuated markedly in real life groups. Time sampling in that case would have been unwise.

Another problem one may encounter while using time sampling is that the psychological meaningfulness of the behavior being observed may be destroyed. For example, rating a group's reactions to a given act of a leader may not give a meaningful picture that is, it may be atomistic or incomplete. When one observes the complete sequence of events in a group's reaction to a leader's act before recording anything, however, it is possible to rate or otherwise record the behavior in a way that includes all of the group's reactions and all of the implications of these actions. This is an aspect of the context problem discussed earlier. Here again the theory of the study

helps one to make decisions. If one simply needed to know the frequency of oral statements which followed each act of the leader, time sampling could be used with little fear that the data would be inaccurate. If, however, one needed to know how constructive the behavior of the group was after a given act of the leader, one would find that the whole sequence of events subsequent to the act would always need to be observed and that a time sampling procedure might cut off observation in the middle of the psychologically meaningful data needed.

To summarize, there are certain criteria which will help us decide whether we should use time sampling in observation. It should be used only when it is apparent that *all* of the behavior which is relevant to the study cannot be recorded. The theory of the study can guide us in selecting the best sampling procedure so that it does not destroy the reliability or the psychological meaningfulness of the material being noted.

TRAINING OBSERVERS

The use of group observers means that we are using people as measuring instruments. A good measuring instrument is one which will accurately measure at various times what it is supposed to measure. If a person is to become an observing instrument, he must be trained to see what it is required that he see. This may be a simple or a complex task, depending upon the data needed. It is an important phase of the preparation of a study using group observers and cannot receive too much attention. A well developed observation schedule will be only as good as the skill of the persons who are asked to use it.

The steps in training observers are fairly obvious. They are worth reviewing here to point out practices which have been found successful and problems which one can expect to meet when observers are being trained.

1. The training process typically begins with a description of the theory and the purposes of the study. This is important since it serves to provide the observers with a reason and thus a motivation for doing well. More important, however, is the fact that it explains why the observation schedule is constructed as it is. Once

the observers are thoroughly familiar with the theory and purposes of a study, the trainer can communicate with neophyte observers concerning the inner workings of the processes being observed the boundary conditions surrounding certain categories and the need for perceiving certain familiar behavior in an unusual fashion

2 If possible, it is helpful if the trainees attempt observation without the aid of a refined observation schedule This means that they would watch a situation comparable to one they would see in the actual study and would attempt to identify as many of the relevant behaviors as they could The discussion among the observers of the events and behaviors they have seen accomplishes several things First, it makes the observers aware that behavior can be seen which might not ordinarily be noticed unless they were looking for it They become sensitive to the fact that the behavior, or other variables which they will be asked to describe, can be found if they are set to watch for them Secondly, this unskilled performance on the observer's part will reveal the need for operating in the future in terms of a carefully defined set of categories The need for careful agreement among them as to what to watch for and how to record the behavior will become apparent as they discover that there are disagreements among them as to what happened as to the significance of certain events as to what to call certain activities and so on

3 The observers are ready, now, for a more refined instrument and the observation schedule may now be introduced We are assuming that the process of developing it has already been completed and that the schedule is prepared for use In our experience the observers are always struck by the complexity of the observation instrument, no matter how simple it may be A set of observation categories or rating scales will look more complicated to the observer at first glance than it turns out to be in practice If one counts on this proper reassurances and practice procedures can be planned

Each of the items in the observation schedule is explained and questions are answered In most cases it is helpful if each observer is provided with an instruction booklet which describes the purpose of the study, the purpose of each item to be observed, cues which may be used for each category, the solution for certain marginal cases the nature of an adequate notation sampling instructions other procedural instructions, etc Obviously, the more the observer is required to make inferences or interpretations of given phe

nomena the more specific and detailed must such instructions be to achieve clarity of understanding

4 The observers are now ready for trial use of the form. It is usually adequate to make the first attempts while they observe a group role playing the type of behavior they will be observing. The value of such a procedure is that any trainee may stop the role playing whenever he has a question concerning the proper coding for a given event.

5 This try out is followed by extensive discussion of the experience. The trainees will have had many problems in selecting proper categories, sampling, keeping up with the events, etc. These questions are answered by discussion and further practice. It often happens that new observers have suggestions for revising the observation schedule which make considerable improvement in its efficiency as a measuring instrument.

6 Whenever the study allows, it is advisable that the observers have an opportunity to make a pilot run on a group like those they will be asked to observe. This will assure that uniform practices are developed for such problems as introducing the observer, etiquette of the observer while watching the group, and any additional problems which arise in the actual test situation and which were not foreseen in the role playing and schedule development phases.

7 Either the pilot run or a later trial will serve to provide data for determining whether the observers are doing a comparable job. Whenever possible, it is best of course, that the research worker be assured that his team of observers are reliable before the actual data gathering begins.

In general, one can expect that the observers will have the greatest problems on those categories which require integration or collation of complex phenomena. They will have the least difficulty, in contrast, with those events which are simple objective occurrences which require little insight or sensitivity on the part of the observers. Studies of coder reliability (7) have found that there is most disagreement on data which are complicated and demand much inference. Since an observer functions as a highly trained coder, it is quite likely that he will have similar problems.

This suggests that the skills required of an observer cannot be performed by all persons equally well, quite aside from the academic training they may have had. One may find that a group of observers have comparable ability to understand the phenomena involved in

a study and to discuss them intelligently but that some of them may not be able to "see" these things in a group of interacting people. A study by Luszki (11) provides some insights which are relevant. She found that there is a positive relation between sensitivity to the feelings and emotions of others (empathic ability) and the ability to "see" what is happening in a social situation. Persons with good empathic ability (1) are better able to see what is happening in the role performance of others, (2) have good personal adjustment, as measured by the instruments used in the study, (3) have insights into themselves which are similar to the evaluations made of them by others, (4) have more stable, positive, and secure feelings about the self and somewhat favorable perceptions of others, (5) have a more consistent and more favorable perception by others.

One may assume, then, that all persons will not do equally well as observers and some testing of competence may force the experimenter to retire less capable ones from the research.

As we have already stated, the observers often make valuable proposals for improving the observation procedures. Thus, it is important that they have the opportunity to participate in making suggestions. In some cases, in fact, the training may be more effective if the observers participate in all stages of constructing the observation schedule. Bavelas (3) trained two groups of observers. One was given training in the use of a prepared observation schedule in a manner similar to that described earlier; the other group participated in the construction of the categories from the very beginning. In the light of its knowledge of the purposes of the study, the group was able to determine the nature of the categories, the definition of each, and the rules for their use. Bavelas reports, in this unpublished study, that those observers who *participated* in the development of the observation scheme were trained more quickly (as measured by the length of time it took to achieve good reliability among themselves) than were those observers who were *told* what the categories meant and how they were to be used, even though the latter group had more instructional time given to them.

unique problems which arise in connection with observational instruments

Validity

The problem of the validity of measures of social interaction has for a number of reasons often been slighted. A brief discussion of these reasons may illuminate the problems which arise.

One way of phrasing the problem is to say that the validity of a measure varies with the degree of relation it has to an external independent criterion. This, as Stouffer points out (13), is a problem of prediction. As was indicated in Chapter 6, there are actually two kinds of prediction situations. In the one case there is a generally accepted measure of the variable in question and the validity of the new measure being developed is determined by its relationship to this other, independent measure. One of the reasons for neglecting the validity of measures of social observation has been the lack of such external criterion measures. It is difficult, for example, to think of an external criterion of a category such as 'Shows dependence'.

The second kind of prediction situation might be described as the prediction of a relation which is derived from a well formulated theoretical system. In this case, if the proposed measure makes differential predictions concerning the behavior of people in a social situation and these predictions are confirmed in an experiment, one could argue that the validity of the measure had been demonstrated. Let us take a simple illustration. Suppose that we had available to us a valid measure of feelings of hostility toward authority figures. Suppose, further, that our theory led us to state the conditions under which hostile behavior in a social situation would develop as a function of these feelings. And suppose that on the basis of the test scores, we organized two groups: one composed of hostile and the other non-hostile people, and that we observed them in such a situation. If, then, our observer measure of amount-of-hostility shown differentiated between the two groups, we could say that it was a valid measure of hostile behavior.

Sometimes the kind of prediction one makes is even less direct than in this illustration. For example, one might predict that two groups of people differing in the amount of hostility shown in a

social situation might differ in their receptivity to influence attempts. If, then, the measure of hostility shown predicted the amount of change which actually occurred, one might argue that the validity of the measure had been established. The statement made earlier, that the theoretical formulation must be precise, ought to be underscored. In other words, the conclusion concerning the validity of a measure in the last instance given depends on whether or not this relation might be predicted.

Sometimes experimenters are satisfied with the validity of an instrument when it relates in a significant way to some other variable. This is a pure case of being satisfied with the validity of an instrument on the basis of its predictive efficiency—it is what Cronbach (4) refers to as *empirical validity*, in contrast to *logical validity*. Logical validity asks 'Is this measure measuring what it is supposed to measure?

One of the difficulties involved in establishing external criteria in social observation research is that social psychological theory often prevents one's having faith in what might superficially appear to be a satisfactory external criterion. Suppose, for example, that observers were asked to note the extent to which the group members were satisfied with the leader. It might superficially appear that one way of checking on the validity of this observer rating would be to ask the members of the group themselves how satisfied they were with the behavior of the leader. There are good reasons, however, for thinking that in many circumstances it would be impossible to accept the report of the members as a validity check. There might, for example, be strong social pressures against reporting dissatisfaction with the leader. Thus in many social psychological experiments, the relation between observer ratings and external criteria is a theoretical and not a methodological problem.

Most of the validity problems with social observation schedules arise in connection with category or rating systems requiring a great deal of inference on the part of observers. There are many category systems which require little inference and the validity is established by the content. In these cases the validity question is not important or critical. For example, the problem solving category system described earlier which has been utilized in describing the problem solving behavior of a group has in it categories such as 'Proposes solutions' 'Gives information'. These categories have

what Guttman calls internal validity (13) In these category systems, the validity is established on the basis of acceptability There is common acceptance, for example, that a question is a question or that an item of information is an item of information Validity based on this common acceptance requires more than anything else a precise definition of the cues to be used in assigning a response to a category system

Another reason for neglect of validity problems is the nature of the variables with which social psychologists are concerned They are frequently complex highly inferential variables for which adequate external criteria are rarely available The social psychologists with their social observation systems are thus often in much the same situation as are clinicians with their complex, highly inferential diagnostic categories which are based on projective test results

In spite of these admittedly sizable obstacles to the establishment of the validity of an observational system it is clear that researchers in this area have not given the problem of validity as much consideration as it should have received Many of the techniques mentioned in Chapter 6 for establishing the validity for a measuring instrument are appropriate for use in observational systems The reader is referred to these for precise designs which might meet his validity problems

Reliability

The reliability of observational instruments has been much more a matter of concern to investigators of social behavior than has validity For purposes of exposition it may be useful to make a distinction that is not often made In assessing the reliability of a system of social observation it is necessary to differentiate the reliability of the behavior being observed from the reliability of the categorization or rating which is made of that behavior In other words the reliability of the observer and of the behavior are separate problems It is clear, of course, from this distinction that the consistency of behavior is a substantive problem, whereas the consistency of an observer is a methodological problem Once the consistency of an observer has been established, it becomes possible to tackle the problem of the consistency of the behavior

The task of assessing degree of agreement among observers in

the case of a category system, breaks down into at least two separate reliability problems. The first is the extent of agreement among observers with respect to the number of units coded. This is essentially the extent to which the observers agree as to the boundaries of the unit which is to be categorized. Guetzkow (7) has pointed out, as has Bales (2), that unreliability in this area is an important factor affecting the reliability of categorization itself. It seems clear that until there is rater agreement as to the boundaries of units to be categorized, there is little purpose in assessing rater agreement on categorizing. The second principal task in assessing reliability is to determine the extent to which observers agree on the category or rating they assign to a specific unit of behavior.

The most frequently used statistic in appraising degree of agreement between observers has been the correlation coefficient. This is especially useful, provided the assumptions underlying its use can be met, because the extent of agreement that is being obtained can be evaluated in terms of a fixed standard. It is a separate question, however, whether in a given case one needs a correlation of 0.7, 0.8, or 0.9 to be certain that one's degree of agreement is satisfactory. It is apparent that the question immediately becomes satisfactory for what? In other words, the experimenter or investigator must ask himself the purpose of his observational scores. His theory will indicate the extent to which large or small differences are to be expected. It is a truism that where fine differentiation is necessary the scores must be more reliable than where gross differences can be expected. It is impossible to state categorically that observational scores should be at such and such a level of reliability to be useful, for the usefulness of a score depends on the use to which it is to be put.

The second principal statistical device which has been utilized in characterizing amount of agreement between observers has been a percent agreement score. This is essentially a matter of computing the percentage of the total number of items which were classified in the same way by the two observers or by all the observers combined. Sometimes it is useful to modify the equation slightly by having as the numerator the percentage of items on which there was agreement, and in the denominator the sum of the items on which there was agreement and disagreement. It is necessary with either of these equations to have some fixed number of items sub

what Guttman calls internal validity (13) In these category systems, the validity is established on the basis of acceptability There is common acceptance, for example, that a question is a question or that an item of information is an item of information Validity based on this common acceptance requires more than anything else a precise definition of the cues to be used in assigning a response to a category system

Another reason for neglect of validity problems is the nature of the variables with which social psychologists are concerned They are frequently complex, highly inferential variables for which adequate external criteria are rarely available The social psychologists with their social observation systems are thus often in much the same situation as are clinicians with their complex, highly inferential diagnostic categories which are based on projective test results

In spite of these admittedly sizable obstacles to the establishment of the validity of an observational system, it is clear that researchers in this area have not given the problem of validity as much consideration as it should have received Many of the techniques mentioned in Chapter 6 for establishing the validity for a measuring instrument are appropriate for use in observational systems The reader is referred to these for precise designs which might meet his validity problems

Reliability

The reliability of observational instruments has been much more a matter of concern to investigators of social behavior than has validity For purposes of exposition it may be useful to make a distinction that is not often made In assessing the reliability of a system of social observation, it is necessary to differentiate the reliability of the behavior being observed from the reliability of the categorization or rating which is made of that behavior In other words, the reliability of the observer and of the behavior are separate problems It is clear, of course, from this distinction that the consistency of behavior is a substantive problem, whereas the consistency of an observer is a methodological problem Once the consistency of an observer has been established, it becomes possible to tackle the problem of the consistency of the behavior

The task of assessing degree of agreement among observers in

determined for each set of conditions, and the degree of reliability desired by the investigator will then determine the number of judges he will need. Using more judges to increase the reliability of observer scores does not seem to us to be a satisfactory substitute for improved precision in definition of the category or the trait to be rated, nor should it be used in lieu of more precise elaboration of cues to be used by the observers in assigning behavior to categories, or instead of adequate training of observers.

In some cases the investigator may find it useful to combine categories to obtain a satisfactory level of rater agreement. This is true because often the lines of demarcation are fine and the distinction is difficult to make. Where it is possible, without introducing conceptual confusion, the investigator may expediently combine categories between which observers have difficulty discriminating.

THE RELATION OF THE OBSERVER TO THE GROUP BEING OBSERVED

One of the most frequent questions addressed to the investigator who reports a study which used observers is "Did the presence of the observers influence the behavior of the group?" Usually he must answer that he has no evidence that the observers influenced the results in any way, but that they might have. It is quite conceivable that the presence of an observer may be an important variable. This depends upon the nature of the group, the type of observation, the nature of the group's activity, the nature of the variables being observed, and a number of other things. Arsenian (1) found that the simple presence of an adult sitting near the door seemed to lend assurance to a group of nursery school children. Yet the presence of observers was a threat to young boys at a summer camp, according to Polansky (12). The influence of observers is a methodological problem which needs more careful study.

Deutsch (5) found that the members of small groups which met frequently over a period of three weeks were aware of the presence of observers at the beginning of their work together but had become almost oblivious to them by the end of that period. Half of these groups were constructed in such a way that the members cooperated with one another to achieve a common goal; the rest of the groups

jected to classification Guetzkow (7) has provided statistical devices for evaluating the level of agreement

Bales (2) illustrates the use of the χ^2 square to evaluate degree of observer agreement. He points out that his use of χ^2 square is different from the more conventional applications since his use applies it to a situation which does not represent random sampling. The principal advantage of χ^2 square is that it does not require the assumptions of the usual parametric techniques.

It seems useful to point out that the investigator ought to be concerned with the reliability of the measure *actually used* in the analysis of the data. For example, it is a matter of *relative* unimportance whether observers agree with the respect to the number of units of behavior assigned to a specific individual if the score with which the investigator is concerned is the number of units in each category made by the group as a whole. In that case the reliability of individual scores is not of concern as long as high group reliability is present. Another illustration is the reliability of the categorization of each unit as opposed to the reliability of a category score for an individual. That is, the agreement between observers as to the percentage of responses categorized in each of the categories in the system may be very high even though the observers show relatively low agreement on the categorization of each item. This condition occurs, of course, only when the errors of categorization are fairly random. The point is, however, that the investigator must ask himself what score he is going to use and must then assess the reliability of that score.

The section concerned with problems involved in the organization of a category system emphasized certain decisions which must be made, many of them with considerations of reliability in mind. The immediately preceding section concerned with the training of observers also emphasized certain problems with the ultimate reliability of the scores in view. Detailed procedures, statistical and design wise, for determining the reliability of an observer rating or observer categorization system can be found in most textbooks in statistics. However, let us make two final remarks.

It has been demonstrated that the reliability of ratings increases with the number of judges. This, of course, depends in part upon the particular trait and the manner of making the ratings, but the generalization seems to hold. In any case, the reliability should be

served, the group members will perceive them as such. For most situations that is probably best.

SUMMARY

The great increase in the past few years in the use of observation methods has been accompanied by an increase in methodological sophistication. Although there is still much methodological research to be done, there is a considerable body of experience to guide the experimenter.

There are two principal types of observation instruments: category systems and rating scales. A category system consists of one or more categories or statements describing a class of phenomena into which observed behavior may be coded. Category systems differ on the following dimensions: *Exhaustiveness*—how much of the total observable behavior is classified into a defined category, *Inference*—the amount of inference (concerning motives, feelings, etc.) required from the observers, *Number of dimensions*—the number of different frames of reference required by the system, *Discrete vs. continuous*—the extent to which the categories can be ordered on some continuum. There is no simple answer to such questions as: How exhaustive should the system be? How much inference should I require? The principal factor influencing the decisions in these cases should be the theoretical framework guiding the research. The second set of considerations affecting the decisions should be those concerned with the competence or trainability of the available observers, since decisions in these areas have important bearing on reliability and validity of observer scores.

Rating scales have been used to describe the behavior of individuals as well as the activities of an entire group. They may be used to record behavior at frequent intervals throughout an interaction period, or to assess the nature of an entire social event after it has ended. Rating scales are particularly useful when a large number of factors are under consideration and the notation of the occurrence of behavior relevant to these factors would be an enormous task. They are also especially helpful in the early stages of an investigation as a device for the development of clear definitions.

were in a situation in which the members competed with one another. It is interesting to note that the competitive groups were much more conscious of the observers than were the cooperative groups. Thus, procedural differences within a group may influence the way in which the observers are perceived.

Polansky (12) noted that changes took place in the perception by boy campers of the purpose and role of cruising observers. Early in the summer, research observers were well accepted by the campers, who had been told that the strangers were neutral persons who were interested in learning how they could improve the camp. By the third week, however, they had become the objects of aggression by many of the boys. The observers decided that their role was too ambiguous and represented a threat to these boys, many of whom were rebellious against adult authority. It was felt that the campers were projecting their feelings of hostility toward adults onto the observers. After the observers changed their behavior to become more warm, human, and friendly, they found that they were no longer rejected by the campers.

Naturally, we must be cautious about generalizing from such an experience as this. In a laboratory situation, for example, where the group members are working toward a decision, it would no doubt be quite disruptive for an observer to indicate that he feels warm and friendly toward the participants. In short, in some situations we may want one perception of the observer; in another situation we may need quite a different one.

Bales (2) has used observers in a wide variety of laboratory arrangements. They have sat with the group, or behind a one way screen with the group aware they are there, or behind the screen with the group wondering whether they are there. He has found no difference in the behavior of the groups which could be attributed to the influence of these various positions of the observers.

Thus, it appears that group observers sometimes influence a social situation. When in doubt, it is wisest to explain the function of the observers to the group in some way that does not spoil any necessary naiveté of the subjects. The observers will do best, barring any special conditions such as those met by Polansky (12), by showing all of the external signs of a piece of furniture. If they behave like a person who is having no reactions to the events being ob-

- 3 Bavelas, A *A study comparing two methods of observer training* Unpublished manuscript
- 4 Cronbach, L J Response sets and test validity *Educ Psychol Measmt*, 1946, 6, 475-494
- 5 Deutsch, M An experimental study of the effects of cooperation and competition upon group process *Hum Relat*, 1949 2, 199-231
- 6 Fouriez, N, Hutt, M, and Guetzkow, H Measurement of self oriented needs in discussion groups *J Abnorm Soc Psychol*, 1950 45, 682-690
- 7 Guetzkow, H Unitizing and categorizing problems in coding qualitative data *J Clin Psychol*, 1950 6, 47-58
- 8 Heyns, R W, and Berkowitz, L *Interaction category systems some problems in development and use of social observation instruments* Mimeographed Conference Research Project, Michigan, 1949
- 9 Jack, L M An experimental study of ascendant behavior in pre school children *Univ Iowa Stud Child Welfare*, 1934 9, No 3 7-65
- 10 Lippitt, R, and Zander, A *Observation and interview methods for the leadership training study* New York Boy Scouts of America 2 Park Ave., 1943 Mimeographed
- 11 Luszki, M B *Empathic ability and social perception* Unpublished Ph.D thesis, Univ of Mich., 1950
- 12 Polansky N, Freeman, W Horowitz M, Irwin, L Papania, N, Rapaport, D, and Whaley F Problems of interpersonal relations in research on groups *Hum Relat*, 1949 2 281-291
- 13 Stouffer, S Guttman L Suchman E A Lazarsfeld P F Star S A and Clausen J A *Measurement and prediction Studies in social psychology in World War II*, 4 Princeton N J Princeton Univ Press 1950
- 14 Zander, A *Third annual meeting of the World Federation for Mental Health A study of an international conference* London World Federation for Mental Health, 1950

of dimensions and the specification of the relevant cues. Under ideal conditions, rating scales should provide data which are strictly comparable to those obtained with a category system with continuous categories.

In the construction of any observation system, the investigator must make decisions concerning the frame(s) of reference to be used by the observers. This involves deciding (1) how much of the context of an act is to be taken into account in the coding or rating, (2) whether the judgments as to the intent of the actor are to affect the classification of the act, (3) whether the consequences of the act are to affect the observers' scoring. Failure to decide these issues is a common source of lack of agreement among observers. Another decision area has to do with the size of the unit to be categorized or rated. Units may vary from single acts to total meetings. Failure to specify clearly the unit to be assessed also affects interobserver agreement. Finally, there is the problem of deciding whether to record all of the relevant behavior in a given interaction period or to sample the behavior. In general, there is considerable risk involved in sampling unless one has adequate theory to guide the sample selection.

The use of observers means the use of people as measuring instruments. This requires the careful calibration of personnel. The training process requires that the observers (1) be familiar with the theoretical framework of the purposes of the investigation, (2) have experiences in which they become sensitive to the dimensions under consideration, (3) have extensive training experience with the proposed observation schedule, including a trial run, preferably with a group like that which they will be observing.

BIBLIOGRAPHY

- 1 Arsenian J. M. Young children in an insecure situation. *J. Abnorm. Soc. Psychol.*, 1943, 38, 225-249.
- 2 Bales R. F. *Interaction process analysis*. Cambridge: Addison-Wesley, 1950.

PART IV

The Analysis of Data

There are few, if any, problems of analysis which are peculiar to research in social behavior. There are, however, three general problems which, although not unique to this area, assume major importance for researchers in this field. The three chapters in this section deal with these three major problems.

Much of the data in social research is collected in what may be called qualitative form. The techniques of translating such data to a form which can be subjected to more rigorous analysis are presented in the first chapter of the section.

Problems of scale construction are prevalent in almost any empirical field. It seems, however, that the solutions to these problems which have proved adequate in other areas do not handle adequately most of the scaling problems encountered in social research. The

Analysis of Qualitative Material

Dorwin P. Cartwright

One of the basic skills required of the social psychologist is that of analyzing symbolic or "qualitative" material. A remarkably large portion of modern social psychological research consists in classifying, ordering, quantifying, and interpreting the verbal and other symbolic products of individuals and groups of people. In this chapter we shall consider some of the kinds of materials which may be analyzed systematically, the major principles involved in converting symbolic "phenomena" into scientific "data," some criteria useful in guiding decisions that must be made in constructing the system of categorization, and some practices found to be helpful in the actual process of categorizing symbolic materials.

The special problems associated with collecting and recording symbolic behavior and those involved in the statistical manipulation of the processed data are treated in other chapters of this volume. Although it is convenient to discuss these separate topics in separate chapters, it is important to realize that, in practice, decisions about the analysis of these materials cannot be made apart from the total plans for the collection and statistical treatment of the data. The ways in which the data are collected will set up severe limitations on the types of analysis which will be feasible. And, in turn, the

second chapter in this section presents an approach to scaling which is perhaps more adaptable to this field

Testing the statistical significance of research findings has always been a problem where the findings are based on relatively few cases and the distributions are highly irregular. This situation is a relatively common one in much of the research done in this field with which we are concerned. The development of "distribution free" statistical techniques may solve some of these problems. The last chapter in this section describes the more important of these techniques

The systematic description of these phenomena by social scientists involves the recording of these symbolic products in an orderly fashion, classifying or categorizing them and determining their quantitative incidence and interrelations. If these procedures are carried out in a proper way, objective and general statements may be made about them.

Qualitative Material Created by Social Psychological Research

Many of the techniques of research developed by social psychologists have as their end product verbal or other symbolic material. The research interview is a principal example of this technique. Here the researcher, by asking questions stimulates verbal behavior which he hopes will provide indicators of certain characteristics of the individual or of his relationships with others. Such variants of the interview as projective tests, stimulated themes, life histories, and the like are of a similar nature. Experiments in the laboratory and in the field also produce materials that must be submitted to systematic analysis.

In research where the symbolic material is specifically stimulated, this material is usually taken to be indicative of something beyond itself. A particular statement, for example given by a respondent in an interview has significance to the researcher because it may be taken to indicate the presence of a certain attitude, value, cognitive structure, or the like. The qualitative analysis of such statements therefore, must proceed in a way that will make it possible to describe clearly to other scientists how the conversion was made from a particular range of qualitative phenomena to a specific genotype or hypothetical construct.

Categorizing Qualitative Materials

When the social psychologist has obtained a set of qualitative materials either from records of natural social phenomena or as products stimulated by a research project he will want to classify the content into appropriate categories so that he can describe it in an orderly way. This process of classification into categories is commonly known as content analysis or coding. The former term is more frequently used in reference to qualitative materials re-

kind of analysis done will limit the manner of statistical treatment that will be permissible and effective

IMPORTANCE OF TREATING SYMBOLIC PRODUCTS SCIENTIFICALLY

Social psychologists are concerned with the analysis of qualitative material for two primary reasons. The proper subject matter of social psychology consists, in large measure, of verbal and other symbolic behavior as it is found in society. Methods must be devised to treat this behavior analytically. But social psychologists do not confine themselves simply to recording and describing symbolic behavior as it is found in 'real life', they also construct situations designed to elicit symbolic behavior under more controlled conditions. In a certain sense, they "create" symbolic materials so that they may analyze them in keeping with the objectives laid down in the design of these contrived situations.

Qualitative Material as Natural Phenomena

When one stops to think of it, it is really surprising how much of the subject matter of social psychology is in the form of verbal behavior. The formation and transmission of group standards, values, attitudes, and skills are accomplished largely by means of verbal communication. Education in the schools, in the home, in business, in the neighborhood, and through the mass media is brought about by the transmission of information and by the exercise of controls which are largely mediated through written or spoken words. If one is concerned with problems of social organization, the situation is similar. Supervision, management, coordination, and the exertion of influence are principally matters of verbal interaction. Social and political conflicts, although often stemming from divergent economic interests and power, cannot be fully understood without studying the words employed in the interaction of conflicting groups, and the process of mediation consists largely of talking things out. The work of the world, and its entertainment too, is in no small measure mediated by verbal and other symbolic behavior.

The systematic description of these phenomena by social scientists involves the recording of these symbolic products in an orderly fashion, classifying or categorizing them, and determining their quantitative incidence and interrelations. If these procedures are carried out in a proper way, objective and general statements may be made about them.

Qualitative Material Created by Social-psychological Research

Many of the techniques of research developed by social psychologists have as their end product verbal or other symbolic material. The research interview is a principal example of this technique. Here the researcher, by asking questions, stimulates verbal behavior which he hopes will provide indicators of certain characteristics of the individual or of his relationships with others. Such variants of the interview as projective tests, stimulated themes, life histories, and the like are of a similar nature. Experiments in the laboratory and in the field also produce materials that must be submitted to systematic analysis.

In research where the symbolic material is specifically stimulated, this material is usually taken to be indicative of something beyond itself. A particular statement, for example, given by a respondent in an interview has significance to the researcher because it may be taken to indicate the presence of a certain attitude, value, cognitive structure, or the like. The qualitative analysis of such statements, therefore, must proceed in a way that will make it possible to describe clearly to other scientists how the conversion was made from a particular range of qualitative phenomena to a specific genotype or hypothetical construct.

Categorizing Qualitative Materials

When the social psychologist has obtained a set of qualitative materials, either from records of natural social phenomena or as products stimulated by a research project, he will want to classify the content into appropriate categories so that he can describe it in an orderly way. This process of classification into categories is commonly known as "content analysis" or "coding." The former term is more frequently used in reference to qualitative materials re-

corded from nature, the latter is more commonly employed in the analysis of materials created by the research. Coding is used especially to refer to the process whereby answers to interviews are categorized. However, no universally accepted usage has emerged to distinguish one term from the other.

In an excellent discussion of the field of content analysis as it has developed in communication research, Berelson proposes the following definition. Content analysis is a research technique for the objective, systematic, and quantitative description of the manifest content of communication (6, p. 18). This is a satisfactory definition if it is interpreted liberally. Communication should be thought of as any linguistic expression, and the restriction to "manifest content" should be removed. With these modifications, we have an adequate designation of all the kinds of analysis of qualitative materials of interest to social psychologists. In the subsequent discussion we propose to use the terms content analysis and 'coding' interchangeably to refer to the objective, systematic, and quantitative description of any symbolic behavior.

Uses Made of Content Analysis

The most detailed summary of the many uses of content analysis is that of Berelson (6), who has developed a system of classification resulting in a listing of sixteen uses of content analysis of verbal material. Although there are several alternative ways in which the work in the field could be classified, Berelson's listing is quite satisfactory. We reproduce it here, with some of the studies cited by him, in the interest of standardizing terminology. For a comprehensive bibliography of publications dealing with content analysis the reader is encouraged to refer to Berelson's book.

Three broad approaches to the analysis of symbolic materials are designated by Berelson. In the first, the researcher is interested primarily in the characteristics of the content itself. In the second he tries to make valid inferences from the nature of the content to characteristics of the producers of the content or of its causes. In the third, he interprets the content so as to reveal something about the nature of its audience or of its effects. Any single study may or may not adopt more than one of these approaches.

CHARACTERISTICS OF CONTENT Interest in the first approach

will lead one to focus either on the substantive nature or upon the form of the content. Berelson lists six uses which are concerned primarily with the substantive characteristics of the symbolic materials. In the first two of these, comparisons are made among materials produced at different points in time. In the next two materials coming from different sources are compared. In the fifth use, the observed substance of communication content is evaluated against standards adopted by the investigator. And under the sixth heading, Berelson simply points out that the substantive characteristics of symbolic behavior are often analyzed by researchers investigating reactions under controlled conditions.

To describe trends in communication content Many investigations have been undertaken to determine changes in content over periods of time. If one is to establish the nature of such trends in communication content, it is necessary to employ comparable methods for sampling the total flow of communication at successive points in time and to use the same system of classification throughout. A fairly typical example of a trend study is the analysis by Yakobson and Lasswell (52) of May Day slogans in the Soviet Union. They found, for example, that May Day slogans changed over a period of years from employing universal revolutionary symbols to 'national' ones. Ojemann (40) studied a considerably different type of content. He recorded trends in articles on child development appearing in the *Ladies' Home Journal* and *Good Housekeeping* and was able to show that articles of this sort were in the early part of the century, much less frequently based upon scientific authority than they were by 1940. Still another kind of trend study is that in which public opinion is measured by sample surveys. Here, instead of relying upon the recording of natural phenomena to reveal trends, the social scientist repeatedly applies the same questions to comparable (sometimes identical) samples of the population in order to detect changes of opinion. Cantril's study (11) of American attitudes toward international affairs just before and after Pearl Harbor, and Cartwright's research (13) on attitudes toward the government's inflation-control program throughout World War II are examples of this use.

To trace the development of scholarship This use of content analysis is essentially the same as the one above. It is mentioned separately because a substantial amount of research has been done

specifically to detect trends in the publications of scholarly and scientific journals. A good illustration of this kind is Allport and Bruner's study (9) of the topics of research in psychology over a period of fifty years.

To disclose international differences in communication content With increasing interest in problems of international relations, social scientists are coming more frequently to study the systematic differences that exist among countries in the content of their major media of communication. Two studies comparing Germany and United States may be cited. Herbert Lewin (34) made a comparable categorization of the literature of the Hitler Youth and the Boy Scouts of America in terms of their goals and justifications. McGranahan and Wayne (35) compared the major themes of the most popular dramas appearing in Germany and America in the years 1927 and 1910. In both these studies, substantial and rather similar differences were found between the two countries. Other studies comparing different countries have been made in terms of such media as radio, newspapers, and textbooks, and a few comparable interviewing surveys have also been conducted in different countries. There are obviously many difficult problems of sampling and of translation in making such cross national comparisons, but this kind of research has produced some of the most useful data now available for an understanding of national differences.

To compare media or "levels" of communication Students interested in understanding the role of the mass media in molding public opinion have made especial use of this type of content analysis. Lazarsfeld, Berelson, and Gaudet (30), for example, studied differences in partisanship among newspapers, magazines, and radio during the 1940 presidential campaign. They found all three media favoring the Republican side, with the magazines more strongly partisan than the other two. Millspaugh (38) analyzed the role of different Baltimore newspapers in the city's interracial relations by studying the treatment given a Negro accused of murder, before his trial, in the different papers. He found sharp differences between the 'white' and "Negro" papers in the proportion of statements carried which were "helpful," 'destructive,' or 'neutral' to the defendant's case. Treatment of many other controversial subjects by the different media have also been compared.

To construct and apply communication standards Many

studies have been aimed at evaluating the social contributions of the media of communication. Such an evaluation can be made only by comparing actual performance against some sort of standard. Criticisms of the media for being biased, or for transmitting frivolous or trashy content, or for oversimplifying issues, etc., call upon, at least implicitly, certain standards of what the media should be doing. If such standards can be made explicit and precise, it is possible for a content analysis of the communications actually transmitted to provide objective evaluation of the media. Even if a technically competent job of content analysis is done, its acceptability as an evaluation will still depend upon the acceptability of the standards employed. Much of the research in this field has been based upon standards proposed by investigators who appeal to such widely accepted cultural standards as "fairness," "objectivity," or "balance." Thus, Sussman, by assuming that radio has an obligation to give a fair and balanced presentation of every major social group, was able to document charges of bias through a content analysis of about thirty news programs on the major networks during a presidential campaign. She found that "labor was presented as being morally wrong five times as often as it was morally right; on the other hand, it was presented as being strong just as often as it was presented as being weak" (45, p. 210). Another example of evaluating performance against standards is provided by the British Royal Commission of the Press (41). This group set up a list of major facts concerning the first year's progress of the National Coal Board. They then checked the reporting of these "actual events" in various newspapers and revealed very meager coverage of what they regarded as socially important information.

To aid in technical research operations. Under this heading, Berelson has grouped two of the most common uses of content analysis in contemporary social research: the coding of free-answer interviews and the analysis of interaction among people in groups. A thorough discussion of the many uses of interview surveys is found in Chapter 1. The problems involved in the latter use are discussed in detail in Chapter 9.

The next three uses of content analysis have in common a focus upon the form of the content (in contrast to its substantive nature). The first of these is concerned with the analysis of propa-

ganda The second derives from a practical interest in improving the intelligibility of written communications The third has been found most commonly among those interested in the study of literature

To expose propaganda techniques Content analysis of propaganda has quite often been designed to reveal the ways in which the propagandist pursues his objectives of influencing the public Berelson points out that two broad classes of "techniques" have been studied the themes or appeals employed, and the "tricks of the trade" Examples of the former are a study of British and German propaganda in World War I by Lasswell (26) and an analysis by White (48) of the values employed by Hitler and Roosevelt in their speeches just before World War II Lasswell concludes, for example, that the British stressed humanitarian ideals much more than did the Germans, and White found, among other things, that 35 percent of Hitler's 'emphasis units' invoked the value of "strength" in contrast to only 15 percent in Roosevelt's speeches A study by Lee and Lee (31) illustrates research on the second type of technique Here seven 'tricks of the trade' were enumerated in the speeches of Father Coughlin Other studies have investigated more particularly the utilization of emotional content Waples and Berelson (47), for example constructed an index of the incidence of emotional terms in various media during the 1940 presidential campaign and found significantly more emotional content in material dealing with Roosevelt than in that dealing with Willkie

To measure the "readability" of communication materials Early interest in the ability to grade materials on the basis of their difficulty of comprehension was displayed by educators wanting to make textbooks fit the age and mental level of different groups of students Various schemes have been developed for this purpose The one currently enjoying the greatest popularity is that of Flesch (19) According to this scheme, two major components of "readability" are isolated *reading ease*, which is measured by the number of syllables per 100 words and by the average length of sentence, and *human interest*, which is measured by the percentage of 'personal words' and "personal sentences" Interesting problems of validation of this measure have been noted, one critic showing that one such scoring system makes Kurt Koffka easier reading than William James

To discover stylistic features. The stylistic characteristics of literary products have been studied quite extensively through content analysis. Typical examples are the examination by Miles (37) of the ratio of verbs to substantives in poetry since the sixteenth century, the study conducted by Skinner (43) of alliteration in Shakespeare's sonnets, and investigations to resolve questions of disputed authorship and of the correct chronology of a given author's works (21, 53).

PRODUCERS OR CAUSES OF CONTENT. The second major approach to content analysis consists of the attempt to learn something about the nature of the producer or, more generally, the causes of the symbolic material from the characteristics of the material itself. In some situations, where the researcher has access only to the communicated material and cannot study the communicator directly, this method is used as a matter of expediency. In other situations, where a person can be induced to produce symbolic behavior as a response to standard conditions, the characteristics of such behavior are often taken as a very acceptable indication of the person's own characteristics. The following four uses of content analysis illustrate various ways in which social scientists have tried to construct a picture of the communicator from his symbolic products.

To identify the intentions and other characteristics of the communicators. In a number of studies the intentions and attitudes of communicators have been inferred from an analysis of the content of their communications. Unfortunately, most of these studies have not had tests of validity that allow a very good assessment of how successful the inferences as to intentions have been. An illustration of this type of content analysis is the study of Britt and Lowry (8) in which the treatment of A.F.L.-C.I.O. conflict in labor newspapers was analyzed to reveal how closely local leadership was following the official position of the national leadership. This analysis revealed a predominant position of neutrality by the local labor press, which was taken by the authors to indicate a considerable inertia on the part of local leaders with respect to the philosophy of the national organization. Another study, in which validation is virtually impossible, is that of Leites *et al.* (32) of speeches given in 1949 by members of the Soviet elite in celebration of Stalin's birthday. These speeches were analyzed so as to disclose attitudes of the speaker

toward Stalin. Sharp differences were found in the image of Stalin revealed by the old Bolsheviks and by the other speakers.

To determine the psychological state of persons and groups

This use of content analysis has perhaps been of most value to people interested in the study of personality. Clinical interviews, projective tests, life histories, diaries, letters, and other personal documents have been analyzed for this purpose. Allport (1) has summarized well the various techniques and objectives of this type of research. In one study Baldwin (3) recorded the frequencies with which certain themes were contiguous to one another within a sample of letters written by a single person. The clustering of themes found in the letters was taken to indicate principal motivational and ideational clusters in the writer's personality. Another type of content analysis which is designed to reveal something of the emotional adjustment of a client during treatment is the discomfort relief quotient developed by Dollard and Mowrer (17). This quotient is computed by dividing the total number of discomfort words by the number of discomfort plus relief words combined. Suggestive conclusions have been drawn from noting changes in the value of this quotient during the course of treatment and from comparing the value of the quotient for one client and another. Quite a different type of study is that of the United States Strategic Bombing Survey (46) in which captured German civilian mail was analyzed to determine the effects of strategic bombing on civilian morale. The letters were coded in various ways to indicate evidence of low morale, poor health, and anxiety. These indications of demoralization were then statistically related to other characteristics of the writers, such as sex, date of writing, tonnage of bombs dropped on locality where the letter was written, and the like. From the relationships thus established, certain conclusions about the effects of bombing on morale were drawn. Thus, for example, day raids were found to be more demoralizing than night raids, and certain community disruptions were found to be more demoralizing than others. In yet another example, content analysis was used to reveal basic personality orientations toward leadership in a study by Sanford and Rosenstock (42). These authors developed a projective device in the form of cartoon pictures which can be administered in brief interviews on the doorstep. They find that responses can be reliably coded and

that scores obtained in this fashion correlate well with an authoritarian-equalitarian scale.

To detect the existence of propaganda (primarily for legal purposes). During World War II the United States Department of Justice, making use of techniques of content analysis developed by Lasswell (27), introduced into the sedition trials evidence that there were remarkable parallels between allegedly native-fascist propaganda and the propaganda of the Nazis. In the scheme of analysis developed for this purpose, Lasswell subjects the material in question to a number of tests. For example: How parallel is the content of a given channel with that of a known propaganda channel? Is the vocabulary employed the same in certain distinctive features as that used by a known propaganda channel? Does the material persistently distort statements on a common topic in a direction favorable to one side of a controversy?

To obtain political and military intelligence. In times of international crisis, when military or political conditions throw an iron curtain around nations, the needs for intelligence about hostile nations assume practical urgency. During World War II and since, methods have been developed whereby it is hoped that national military and political intentions can be anticipated from analyzing the characteristics of nationally controlled communication materials. The Foreign Broadcast Intelligence Service was engaged in such activities during the last war. Although it is extremely difficult to make at all satisfactory checks of validity on these methods, George (20) concludes from a careful study of the performance of these researchers that they made successful predictions about twice as often as unsuccessful ones. Further evaluation of this method cannot be made at present because much of the current research in this area is not available to social scientists generally for reasons of military security.

AUDIENCE OR EFFECTS OF CONTENT. In the third major approach to content analysis, the material is taken as a basis for inference about characteristics of the audience for whom the content is intended, or about the effects of communication. It should be apparent that an inference from the nature of the content to the nature of its audience is possible only if certain assumptions are made (for example, that the communication correctly reflects audience inter-

est) about the situation in which the content is produced. Often these assumptions have little empirical justification and the inferences about the nature of the audience are worthy of consideration only because direct observation of the audience is not possible. The following three uses of content analysis illustrate this third approach.

To reflect attitudes, interest, and values ("cultural patterns") of population groups In several studies of the content of mass media, it has been assumed that the material communicated through these media express or reflect the prevailing thought and mores of the population at that time. Thus, Hart (22) analyzed the content of popular magazines in the United States over a period of years from 1900 to 1930. He found what he took to be evidence for a decline in the status of religion and an increase in toleration of sexual freedom during this period. The primary basis for these conclusions were changes in the amount of attention given to the topic in the magazines and the incidence of indicators of approving attitudes. Wolfenstein and Leites (50) have drawn certain conclusions about the American culture of today from an analysis of contemporary feature films. In this study it is explicitly assumed that a movie is a sort of national 'day-dream' which reveals something of the emotional life of the movie-going public. One of the more ingenious aspects of this study is the interpretation of American relations between the sexes based upon the prevalence of the "good bad girl" in American movies, a type not so common in films of other countries. In all studies of this type, it is difficult to determine why the investigators assume that the content reflects characteristics of the audience rather than of the producers. Often the assumption seems to be either that the content reflects both equally well, or that it reflects the audience because the producers are for some reason (perhaps by a sort of 'natural selection') attuned to the audience.

To reveal the focus of attention A slightly different assumption underlies inferences to the audience in this use of content analysis. Here it is assumed that there is a more or less approximate correspondence between the content of the mass media and the cognitive content of the audiences exposed to the media (presumably because the media produce the cognitive content). If some item of information or theme is stressed in the media at some place and time, it is assumed that this will be salient in the thinking of the

population Research on reading and listening behavior indicates that this assumption can, at most, be only an approximation and that various people exposed to the same content through the same media may react to it in quite different ways Some examples of this approach may be cited Woodward (51) showed that the percentage of foreign news in American morning newspapers in 1927 rarely exceeded 10 percent, indicating presumably, that American people were only slightly aware of foreign events Festinger Cartwright, *et al* (18), in studying the circumstances under which an anti communist rumor broke out in a local community, showed that in the period preceding the outbreak of the rumor there was a marked increase in the number of column inches devoted to the theme of domestic communism in the newspapers read by the local citizens Although these newspapers did not deal at all with the specific subject of the local rumor, the authors argue that the media set or reflected, an atmosphere favorable to the occurrence of the rumor Arnheim (2) conducted a now classical study of what kind of world is brought to the attention of listeners to soap operas He showed, for example, that the world of the daytime radio serial deals predominantly with themes concerning personal problems rather than public affairs Of 43 such serials, for example, 49 percent dealt with problems of courtship, whereas only 26 percent were concerned with public affairs

To describe attitudinal and behavioral responses to communications Berelson points out that there have been three ways in which content analysis has been used to study the effects of communications The first consists in analyzing materials which were produced in response to some specific communication Lerner's study (33) of published reactions to *The American Soldier* is a good illustration of this type of analysis The second kind of investigation attempts to show empirical relations between the content of a communication and responses to it Thus, Berelson (5) showed that the greater the frequency of certain political arguments in the various media, the larger was the number of people who could recognize the argument On the other hand, he found a much weaker relationship between the frequency of appearance of the argument and acceptance of it Merton (36) also attempted to relate characteristics of the content of media to people's reactions to it in his study of Kate Smith's war bond marathon during World War II He noted

among other things, that fully half of Smith's material stressed the sacrifice theme and, in intensive interviews with people who heard Smith, he traced through the personal meaning of this appeal to listeners. In the third kind of analysis, direct inference is made to the effect of content without any reference to response data themselves. In this way Lasswell and Blumenstock (28) analyzed the themes employed by communists in Chicago in the 1930s and concluded that their propaganda was relatively ineffective because it ran counter to the fundamental values and mores of the citizens to whom it was addressed.

It should be apparent from this summary that content analysis has received serious attention in widely different fields of investigation. Any evaluation of the theoretical or practical significance of this technique must be made, as in the case of any technique, in terms of the specific objectives set up for the research. The work completed to date undoubtedly shows that this technique can be successfully applied to the solution of many significant problems. Only future work can demonstrate its full potential and limitations. It is possible, however, from the experience accumulated to date, to set up certain standards which should be met in the process of analyzing symbolic materials.

Let us turn now to a more detailed examination of the process by which qualitative materials are converted into scientifically acceptable data and examine some of the principles that should govern this operation.

CONVERTING PHENOMENA INTO SCIENTIFIC DATA

The recording of symbolic materials as found in life settings and the stimulation of them in contrived situations provide the analyst only with raw materials. Inspection of such materials may lead a sensitive person to certain insights and conclusions and these may be in a certain sense, 'correct'. But, in the long run both scientific and practical progress require more than sensitive insight (though both can certainly make good use of it). To the extent that investigators cannot communicate to others how their insights are accomplished, the ability to achieve them is retained as private

property of individuals. These conditions would produce, at best, experts and not a body of knowledge.

The objective of content analysis is to convert recorded "raw" phenomena into data which can be treated in essentially a scientific manner so that a body of knowledge may be built up. More specifically, content analysis must be conducted so as (1) to create reproducible or 'objective' data, which (2) are susceptible to measurement and quantitative treatment, (3) have significance for some systematic theory, and (4) may be generalized beyond the specific set of material analyzed.

The Problem of Objectivity

Suppose that an investigator has collected such materials as speeches made by a political candidate, articles appearing in *Pravda* or the *New York Times*, deliberations of the United Nations Security Council, or answers given by respondents in a sample interview study. How should he go about constructing his descriptive statements concerning these materials so that other analysts can verify them independently? Four aspects of this problem of objectivity may be noted.

THE VARIABLES TO BE EMPLOYED IN THE ANALYSIS OUTLINE. Unless there is agreement among investigators about the aspects of the material that are to be described, there can hardly be agreement in the resulting descriptions. To note how many different attributes can be found in the same material, consider the brief quotation from an interview conducted with an industrial worker during World War II.

I'd like to see all trade barriers down after the war. Raw goods should be shared where they're needed. It's money and raw goods and poor living that caused most of this war. We should see that Germany gets it fair this time or we'll have another war. Russia is fighting for her way of life just like we are for ours. England is fighting along with us and Russia to protect the people against fascism—to be free, not slaves. Churchill and Roosevelt and Stalin are great men. They know how the people feel. We can't stay on our side of the pond anymore. The unions taught me that.

Literally scores of attributes can be found in this brief passage. Let us list a few: (1) number of words, (2) percentage of personal pronouns, (3) attitude toward free trade, (4) perceived cause of war, (5) degree of confidence in the Allies, (6) degree of confidence in leaders, (7) attractive traits of leaders, (8) attitude toward isolationism, (9) evidence of previous isolationism, (10) source of influence on attitudes, (11) implied values, (12) inclusiveness of cognitive structure, and (13) degree of approval of Allied war aims.

It should be obvious that many other attributes of this material could be listed and that disagreements about the true nature of the material could easily arise. Objectivity requires, therefore, explicit specification of the variables (sometimes referred to as "dimensions" or "types of attributes") in terms of which descriptions are to be made. This is the first step in constructing the analysis outline (or code). We shall return later to the question of *how* these variables should be selected.

THE CATEGORIES FOR EACH VARIABLE Let us assume now that we have chosen some variable, perhaps one from the list above: confidence in the Allies. There remain many ways in which this variable may be broken down into categories. We might decide to code all interviews (or any appropriate material) so that each has to be placed in one, and only one, of the three following categories: (1) High, (2) Low, (3) Not classifiable in either. It would however be equally possible to use seven categories: (1) Unqualified confidence, (2) Qualified confidence, (3) Confidence and mistrust equally balanced, (4) Qualified mistrust, (5) Unqualified mistrust, (6) Question not asked by interviewer, (7) Question asked, but answer not classifiable in above categories. It should be apparent that if two independent analysts were to code the same material, one using the first set of categories and the other using the second, they would come out with different descriptions of the same phenomena. It should also be clear that many other systems of categorization are possible. Explicit specification of the system of categories used with each variable is essential for reproducible analysis.

THE OPERATIONAL DEFINITION FOR EACH CATEGORY Two independent analysts might agree to analyze the interviews for "confidence in the Allies" and to employ the threefold system of categorization, and yet they still might not agree at all in their actual codings. To be sure that they will agree, they need explicit rules

specifying what features of the content are to be taken as indication that it falls in one category rather than another. A statement of these rules constitutes the operational definition of the category.

In drawing up such an operational definition, it is important to begin by designating the units of analysis that are to be used. There are basically two kinds of units to be specified. The first of these may be called the "recording unit," which is the specific segment of the content that is characterized by placing it in a given category. The second kind or unit is the "context unit," which is "the largest body of content that may be examined in characterizing a recording unit" (6, p. 135). The coder might, for example, count each emotionally loaded word as a recording unit, but refer to an entire paragraph to be sure that he records its correct meaning. In the coding of free answer interviews, the answer to a single question is often taken as the recording unit and a whole block of related questions is employed as the context unit. This procedure is followed because the correct meaning of an answer to a single question may sometimes be appreciated only by reference to what has gone before or what comes after.

A second aspect of the operational definition of a category consists in specifying the indicators which determine whether any given unit should fall within the category. In the example given above, we might have considered the category "high confidence in the Allies" and taken as an indicator the statement "England is fighting along with us and Russia to protect people against fascism." If we were coding a very large number of interviews, we would encounter many other statements which should also be taken to indicate the same "high confidence in the Allies." Thus, a category consists of a range of possible indicators, all of which are given the same label and are therefore handled equivalently in all subsequent treatment of the data. If it were possible to list all the variations of content which indicate a given category, such a list would provide a complete operational definition of the category. Unfortunately, most categories with which social scientists deal cannot be defined in actual practice by an exhaustive listing of indicators.

Instead of attempting to construct a complete list, the analyst will find it more effective to rely upon the ability of a trained person to respond to indicators in a systematic way. To respond systematically, the coder needs a rationale for a given set of equivalent indi-

cators Often this can be conveyed by establishing the "core meaning" or 'ideal type' of a given category and then defining its boundaries through examples of indicators which will be taken to fall on each side of the boundary¹

ADAPTATION OF ANALYSIS OUTLINE TO THE EMPIRICAL CONTENT

The most logically constructed and theoretically elegant scheme of analysis will not produce objective results if it does not in fact fit the material being analyzed Try, for example, to fit the interview quoted above into a system of classification designed to reveal the respondent's stage of psychosexual development Reproducible coding will be possible only when the system of classification is properly adapted to the material being coded

Lazarsfeld and Barton (29) have suggested that in constructing the analysis outline for use with free answer interviews, there are two adaptations to the empirical material which should always be made The first of these they call adaptation to the structure of the situation Thus, in the analysis of "reasons given for a certain behavior it is desirable to 'build up a concrete picture or model of the whole situation to which the reports refer, and then locate the particular report within this 'structural scheme'" (p 159) In constructing a scheme for analyzing reasons women give for buying a specific brand of cosmetics, one would set up variables referring to such things as sources of information, sources of advice, motives related to use of cosmetics technical qualities of cosmetics, anxieties about consequences on health, considerations of expense, etc An excellent example of this approach is the study of men's preferences in suits, coats, and jackets, conducted by the Division of Special Surveys of the United States Department of Agriculture (10) In designing the interview and the scheme of analysis, these investigators obtained detailed data bearing on three general areas (1) circumstances that led to the decision to buy the most recently purchased suit, (2) what men wanted their suits to do for them and (3) the values men brought to bear upon their suit shopping and the means by which they thought they could attain them Several more specific variables were set up in each of these areas

¹ Cartwright and Festinger (12 15) have demonstrated that people do employ categories of judgment whose boundaries may vary in precision and that difficulties of categorization as reflected in decision time increase as the material being classified moves from the core of the category toward the boundary

The second type of adaptation to the empirical material is "adaptation to the respondent's frame of reference" (p. 162). The need for such an adaptation becomes vividly apparent whenever one attempts to apply a classification scheme assuming greater sophistication or differentiation than in fact exists in the respondent's thinking. Cartwright (13) found, for example, in studying popular conceptions of wartime inflation control, that characterizations employed by technical economists could not be used in coding popular descriptions of war finance.

This requirement that we adapt the analysis outline to the respondent's frame of reference has one important consequence often overlooked. Consider its implications for two different techniques of interviewing—the free-answer question and the fixed-alternative question. From the free answer, the researcher obtains verbal material which he must analyze according to his scheme of analysis. When, however, a fixed-alternative question is asked, this scheme of analysis is given with the question, and the respondent is asked to code the answer which he would have given had he been allowed to talk fully. Under these circumstances, if the analysis outline does not fit the respondent's frame of reference, the only alternatives open to the respondent are to refuse to answer or to indicate a categorization which is not accurate. Crutchfield and Gordon (16) have produced convincing documentation of this danger by following up a fixed-alternative question with a series of free-answer questions designed to reveal the respondent's frame of reference.

Difficulties in getting an a priori analysis scheme to fit the verbal materials have sometimes led analysts to abandon efforts to construct an analysis outline *before* studying the content of the material. The results of abandoning these a priori considerations tend to be the construction of an outline which reflects only the superficial or phenotypical similarities and differences among the elements of the content. Experience suggests that it is better procedure to start with an analysis outline and then to adapt it in a self-conscious and orderly fashion so as to make it fit the content being studied. In this way, it is possible to examine systematically the modifications in the a priori scheme which are called for. If these modifications are substantial, one may wish to conclude either that the original outline was inadequately conceived or that the material chosen for

analysis was not in fact appropriate. In either event the analyst modifies his original conceptions in an explicit and self-conscious fashion.

We should consider at this point the attempt that several investigators have made to construct a standard or all-purpose scheme of categories to be used in a wide variety of studies. Some of these schemes are extremely phenotypic, consisting of such categories as positive or negative affect. Others, however, are derived from a more or less developed conceptual system. Examples of these more genotypic standard schemes of analysis are Bales' (4) categories for analyzing interactions in discussion groups and Whites' (49) categories for describing values employed in verbal materials. There can be little argument about the desirability of having standardized schemes of analysis so that different studies can be compared. It is probably no accident, however, that there are as yet relatively few studies conducted by independent investigators which use these schemes. To be satisfactory, the schemes must fit readily both a generally accepted conceptual system and the specific content being employed in each new investigation.

The Problem of Quantification

One of the major reasons for developing an explicit and objective scheme of analysis is that it makes possible quantification and measurement (provided that certain additional requirements are met). After material has been submitted to a scheme of analysis meeting the four requirements of objectivity listed above, it is possible to determine frequencies, establish quantitative relations, and engage generally in many of the operations usually thought of as measurement. The abstract features of measurement theory are discussed in greater detail in Chapter 6. We shall limit our discussion here to certain common problems and practices encountered in most current work employing content analysis.

THE UNIT OF ENUMERATION The quantitative treatment of symbolic materials requires that one specify clearly the unit in terms of which quantification is performed. We shall refer to this as the *unit of enumeration*. In our earlier discussion we referred to another unit—namely, the *recording unit*—as that segment of the content which gets labeled when the analyst codes the content. It is

important to note that these two kinds of units are not necessarily the same. Sometimes—as, for example, when the analyst merely counts the number of recording units which get a certain categorization—the recording unit is exactly the same as the enumeration unit. An illustration in which the two are the same might be the analysis of a speech given by a public official so as to reveal the number of times ‘American military strength’ is employed as an argument for a certain foreign policy. In this case an argument is taken both as the recording unit and the enumeration unit.

But let us consider an example in which the two units are not identical. One might characterize an entire editorial on foreign aid as predominantly favorable or unfavorable and then, for purposes of quantification, count the number of column inches of the editorial. In this case, a column inch would be the unit of enumeration, whereas the editorial as a whole would be the recording unit. Quite different quantitative results may be obtained through use of various units of recording and enumeration. In the latter example, for instance, we might just as legitimately use the sentence as the recording unit so as to be able to count the number of favorable and unfavorable sentences in the editorial. Then we might find that slightly more than half of the recording units in *each* of the editorials are favorable. We should then conclude that 55 percent, let us say, of the sentences are favorable to foreign aid. But if we use the whole editorial as the recording unit and the column inch as the unit of enumeration, we might conclude that 100 percent of the column inches of the editorials is favorable. The units we choose to employ must be determined by the purposes of the total analysis.

In analyzing free answer interviews it is customary to take a single respondent as the unit of enumeration. In this way, quantitative statements are made concerning the number of people who display a given characteristic. Some interview studies, however, have used the recording unit as the unit of enumeration with resulting confusion. Consider an example in which this confusion arises. In an interview each respondent is allowed to give several reasons for taking a given political position—let us say, for favoring a particular candidate. The analyst takes each reason as the recording unit and then uses this unit as the unit of enumeration. Results are reported in terms of the number of times a certain reason appears in the entire collection of interviews without respect to the number of

respondents who mention that reason. The results are now confused. Suppose, for example, that the number of reasons turns out to equal the number of respondents in the study. What can one conclude from this fact? Obviously one can conclude very little, because the same result would have been obtained if each respondent had given one reason or if one fifth of the respondents had given five reasons each. This practice of using some segment of the content rather than the respondent as the unit of enumeration is sometimes defended on the grounds that the analyst is interested in measuring a 'climate of opinion' or 'culture' rather than characteristics of individuals. A convincing logic for this procedure, however, has yet to be developed.

THE SYSTEM OF CATEGORIZATION. Quantification and measurement depend not only upon a unit of enumeration, but also upon the existence of certain systematic relationships among the categories. If the content is to be used as an aid to measurement, and if certain quantitative treatments are to be employed, the categories of each variable must be related to one another in certain definite ways.

Lazarsfeld and Burton (29), in an illuminating discussion of the logic of measurement, make a useful classification of the types of systems of categorization which are possible in coding qualitative materials. These may be referred to as (1) dichotomies, (2) series, and (3) variables.

A system of classification which employs *dichotomies* calls essentially for a judgment of the presence or absence of the attribute in question. Examples of such a scheme are a listing of reasons, or the counting of a certain kind of word, phrase, theme, or value. Here the coder looks at each recording unit and notes either the presence or absence of the attribute under consideration. Reliable coding of this type requires that an explicit judgment of 'presence or absence' be made. Sometimes, under the time pressures of a research project, the coder skims quickly through a body of material looking for certain indicators and noting their presence. The logic of this kind of coding requires, however, that every recording unit for which he does not note the presence of the indicator be taken to mean that the indicator is, in fact, not present. A failure to be sure that an indicator is *not* present often is found to lower the reliability of such coding.

In analyzing many kinds of content, it is desirable to categorize units in terms other than mere presence or absence. For example, one may want to indicate that a statement or attitude has a certain degree of intensity. Thus, instead of categorizing the statement as merely present, one might wish to indicate that it reflects high, medium, or low intensity of feeling. Such a system of categories may be called a *serial*. It orders the categories in such a way that the coded materials may be ranked. This means, for instance, that an indicator categorized as "high" is above those categorized as "medium" and "low," that an indicator coded "medium" is above a "low" one but below a "high" one, and that an indicator coded "low" is below both "medium" and "high" ones. No assumptions are made concerning the location of an absolute zero point. Most scales found in content analysis at the present time are serials. An example of such a scale is the common five category scale of degree of satisfaction, consisting of (1) "very satisfied," (2) "satisfied," (3) "neutral or ambivalent," (4) "dissatisfied," and (5) "very dissatisfied." Another example would be the four category scale of reported frequency of behavior, made up of (1) "always," (2) "usually," (3) "occasionally," and (4) "never." Not all serial scales need to have the superficial appearance of being graduated. Scaling procedures, such as those developed by Guttman, may result in a system of categories which meet requirements of scalability without possessing the obvious and apparent characteristics of a graduated series.

If a system of categories not only establishes a serial order but also designates equal intervals and an absolute zero, it meets the full requirements of a *variable*.² Only a few of the coding schemes employed in content analysis meet the requirements set for a true variable. The most common ones are given in terms of time (such as age of respondent or duration of a radio program), monetary units (such as income, prices, or savings), or units of physical length (such as distance the respondent lives from the public library, or

² The reader will note that the term *variable* has been used in two rather different ways in this chapter. Prior to this section it has referred to the type of attribute being described by a given set of categories. The more restricted meaning of the term refers only to such types of attributes as are categorized by a system of categories meeting special requirements. This double usage of the term seems unavoidable until an acceptable term is adopted for the looser meaning of the word.

respondents who mention that reason. The results are now confused. Suppose, for example, that the number of reasons turns out to equal the number of respondents in the study. What can one conclude from this fact? Obviously one can conclude very little, because the same result would have been obtained if each respondent had given one reason or if one fifth of the respondents had given five reasons each. This practice of using some segment of the content rather than the respondent as the unit of enumeration is sometimes defended on the grounds that the analyst is interested in measuring a 'climate of opinion' or 'culture' rather than characteristics of individuals. A convincing logic for this procedure, however, has yet to be developed.

THE SYSTEM OF CATEGORIZATION Quantification and measurement depend not only upon a unit of enumeration, but also upon the existence of certain systematic relationships among the categories. If the content is to be used as an aid to measurement, and if certain quantitative treatments are to be employed, the categories of each variable must be related to one another in certain definite ways.

Lazarsfeld and Barton (29) in an illuminating discussion of the logic of measurement make a useful classification of the types of systems of categorization which are possible in coding qualitative materials. These may be referred to as (1) dichotomies, (2) series, and (3) variables.

A system of classification which employs *dichotomies* calls essentially for a judgment of the presence or absence of the attribute in question. Examples of such a scheme are a listing of reasons, or the counting of a certain kind of word, phrase, theme, or value. Here the coder looks at each recording unit and notes either the presence or absence of the attribute under consideration. Reliable coding of this type requires that an explicit judgment of 'presence or absence' be made. Sometimes under the time pressures of a research project, the coder skims quickly through a body of material looking for certain indicators and noting their presence. The logic of this kind of coding requires, however, that every recording unit for which he does not note the presence of the indicator be taken to mean that the indicator is, in fact, not present. A failure to be sure that an indicator is *not* present often is found to lower the reliability of such coding.

producers of content, the analyst may want to change his system of categories to produce a better fit. But if he does so, he cannot then make strictly satisfactory quantitative comparisons. This problem is particularly acute in comparisons over long periods of time and in comparisons between widely different cultures, and no fully satisfactory solution has yet been worked out.

When it is possible to set up in quantitative terms certain norms or ideal states, it is then sometimes possible to obtain quantitative measures of the degree of deviation from these norms. Several examples of this type of investigation were mentioned in the first section of this chapter. In order to reveal degree of deviation from a norm, the study must be designed so that the norms and the coded materials are stated in equivalent units. The study of majority and minority Americans in magazine fiction made by Berelson and Salter (7) illustrates one way in which such comparisons may be accomplished. These investigators took as a norm the proportion of the total American population represented by various minority groups and then calculated similar proportions for the appearance of various minority groups in a certain fictional population. Discrepancies could then be stated quantitatively.

In some studies of this type, the norms are stated in terms of some ideal pattern of the coded items. Thus, in an interview, an index of "information about world affairs" might be defined as the percentage of "correct" statements made in answer to certain questions. Here, the ideal state would presumably be 100 percent. Subgroups of the population could then be compared in the amount they deviate, on the average, from this ideal. A somewhat similar approach is illustrated by the coefficient of imbalance developed by Janis and Fadner (23). This coefficient results in a value of zero if the number of favorable statements equals the number of unfavorable ones. Furthermore, a quantitative measurement of deviation from such balance is given by use of the coefficient. If balanced presentation were set up as an ideal (as is often done for mass media in treating controversial subjects), this coefficient could then serve to measure how closely any given producer of content conforms to the ideal.

One ultimate objective of social science research is of course the discovery of causal relations. The fundamental problems in constructing and using any research method bear directly or indi-

column inches of type) It is apparent that these examples do not refer to psychological variables and that the process of coding consists of little more than transcribing answers to the tabulation sheets If it were possible to employ true variables in the analysis of psychological material many mathematical operations not otherwise possible could be performed on the data (for example, treating different points on the scale as ratios)

In treating categorized data quantitatively, the basic operation is that of counting This is true whether the system of categorization is a dichotomy, serial or variable After the material has been categorized the usual procedure is to tabulate the frequencies obtained for each category If the system of categorization is that of a dichotomy the frequency for any given category is usually calculated as a percentage of some total possible frequency Thus, one notes the percentage of all respondents interviewed who mention economic imperialism as a cause of war, or the percentage of all value statements in a speech which appeal to strength If the system of categorization is that of a serial or variable, the frequencies for each category may be noted and such measures as those of central tendency and of dispersion may be calculated

MAJOR REASONS FOR DETERMINING QUANTITATIVE RELATIONS Basically, the social scientist is interested in quantifying symbolic material so that he can compare different sets of material and examine relations in a precise way He may wish to do these things for any of several purposes Let us consider briefly some of the more common of these

There are two basic kinds of questions that are raised in most descriptive studies (1) How do symbolic materials vary over time? and (2) How do materials produced by different sources differ from one another? Many examples of both types of relationships were given in the survey of uses of content analysis presented above In establishing trends over time and in comparing different kinds of materials it is essential that the same system of categories the same operational definitions of the categories and the same units of recording and of enumeration be used in quantifying the materials being compared This requirement is sometimes difficult to meet when the several materials are quite different in content If the frame of reference of the person producing the content changes over time or if different frames of references are used by different

may reveal something of the nature of causation. Although there is ample evidence that people may not have correct insight into the determinants of their behavior, this important source of information should not be ignored.

In the ideal approach to the problem of determining causation, a variety of techniques will be used. It is often possible to combine in the same study the two approaches just mentioned. For example, in studying the war-bond program of the United States government during World War II, Cartwright (13) asked respondents to report why they had bought bonds during a specific drive. In addition, he examined the relationships between reported bond buying and reports of what happened to respondents during the drive. Some people reported that they bought bonds because they had been personally solicited. Analysis of the interviews showed also that people who reported having been personally solicited were much more likely to have bought bonds than people who did not report a solicitation. The agreement between these two types of evidence heightens one's confidence in the conclusion that personal solicitation was a causal determinant of bond buying.

Under the most favorable conditions, causation will be determined by independent manipulation and measurement of the independent variables. The study of demoralization under bombing, cited above (46), approximates this design. Here, quantitative indications of demoralization in written letters were related to intensity of bombing (measured in tonnage dropped) in the community where the letter writer lived. This example shows, incidentally, that manipulation of the independent variable need not necessarily be performed by the researcher himself.

The Problem of Significance

One of the most serious criticisms that can be made of much of the research employing content analysis is that the "findings" have no clear significance for either theory or practice. In reviewing the work in this field, one is struck by the number of studies which apparently have been guided by a sheer fascination with counting. Unfortunately, it is possible for a content analysis to meet all the requirements of objectivity and quantification enumerated above without making any appreciable contribution to theory or practice.

rectly upon this objective. We shall consider here only a few of the special problems involved in using content analysis in determining causation.

The covariation of two attributes is commonly taken to suggest that there might be a causal relation between them. When symbolic material has been analyzed in a quantitative fashion, it lends itself to this type of study. It should be noted that for this purpose the two variables need not necessarily be expressed in the same units. It is possible to assert, for example, that a reduction of income reduces personal optimism without measuring the two things in similar units.

For many purposes a causal analysis is undertaken by looking for covariation of attributes within the same body of content. The contiguity of certain themes in written material was taken by Baldwin (3) as evidence for a functional interdependence. This method is quite common in the analysis of interviews in sample surveys. It may be illustrated by an unpublished study conducted during World War II in which a substantial correlation was shown to exist between expressions of internationalism and statements indicating an equalitarian ideology. This finding was taken to support (but not to prove) the hypothesis that a basic ideology influenced specific attitudes.

Sometimes it is possible to demonstrate covariation between features of coded content and some external variable. In interview surveys this approach is quite common if we assume that such things as reported age, sex, marital status, income, and the like correctly reflect external variables. The index of political predisposition constructed by Lazarsfeld, Berelson, and Gaudet (30) from reported religion, socioeconomic status, and residence illustrates this approach. The investigators showed that certain combinations of these three characteristics of respondents appeared significantly to predispose attitudes toward political candidates. Nearly all attitude and opinion surveys make use of this device of analysis in one fashion or another.

A rather different approach to discovering causation should also be noted. The basic assumption here is that the analyst is sometimes able to discern causation directly from the nature of the content. In interview surveys respondents are often asked to state reasons for their attitudes or behavior. The coding of such reasons

interviews would have been futile for the purposes of the study.

Another way of stating the requirement for a significant content analysis is to assert that the variables of the analysis outline must yield a genotypic rather than phenotypic description of the material. If the content is classified simply according to its superficial similarities and differences, little will be learned of relevance to "pure" theory or "practical" application.

To insist that the variables of the analysis outline must yield genotypic descriptions does not imply, however, that the *coder* must always place the content directly into genotypic categories. Even if a variable is designed to reflect, for example, an attitude toward Negroes, the coder need not necessarily rate responses on an attitude scale ranging, let us say, from "very favorable" to "very unfavorable." It is possible to employ the attitude as a variable and still have the coder categorize responses in more phenotypic terms (perhaps by noting the presence or absence of certain stereotypic characterizations of Negroes). If this latter procedure is employed, it is necessary for the analyst to have some explicit procedure for placing these phenotypic indicators upon the attitude scale. Which of these two procedures will result in more reliable and valid classification will depend largely upon the skill and sophistication of the coder. Usually it is easier to obtain good reliabilities with more phenotypic categories. There is also an advantage in being able to list explicitly the indicators used in rating attitudes. When the content dealt with, however, is complex and subtle, it is often found more economical to employ sophisticated coders who are able to interpret directly the genotypic significance of the material.

The Problem of Generalization

As a rule the content analyst is not interested in limiting his conclusions or findings strictly to the content actually analyzed. Almost invariably he undertakes his specific analysis in order to reveal something about a more general universe of data than just those symbolic materials (produced at a certain place and time) with which he deals. However, generalizations from a limited set of data to a more inclusive universe cannot be made legitimately unless certain conditions are met and certain procedures followed.

In considering the problem of generalization, it is convenient

It is an all too common error to equate 'scientific' with "reliable and quantitative." Unless the findings of a content analysis have implications for some theory, however vaguely formulated, the study can merit serious attention only on the highly tenuous claim that some day the significance of the findings will become apparent.

For this reason, significant content analysis begins with some systematic problem whose solution will be determined by the specific nature of the data resulting from the analysis. This problem may stem either from a desire to extend a theory or conceptual model to some new realm of phenomena or from a need to predict or control events for some practical end. In either case the investigator must have an *a priori* conception of the variables that are relevant to his problem. The purpose of the content analysis is to indicate the presence or absence of these variables in the 'real world,' something about the relative magnitude of the variables, and something about the relations among different variables.

In constructing the analysis outline, it is therefore necessary to compose it of variables which mirror the variables of the researcher's *a priori* conception. Thus, when Hart (22) wanted to use magazine fiction to test the hypothesis that American culture changed during the first quarter of the century in a direction toward greater toleration of sexual freedom, he had to put in his analysis outline variables that would reflect "degree of toleration of sexual freedom." When the Division of Program Surveys (14) conducted an interview study to determine whether people's plans for the use of their war bonds would predispose them to cash bonds for the purchase of consumer goods, the analysis outline used in coding the interviews had to provide dimensions for categorizing "plans for the use of war bonds."

From these examples it should be apparent that the value of a content analysis will depend upon the quality of the *a priori* conceptualization. It will depend, also, upon the adequacy with which this conceptualization gets translated into the variables of the analysis outline. Finally, it will depend upon having data to analyze which are appropriate to the variables of the outline. Had Hart's magazine fiction presented no evidence whatsoever about toleration of sexual freedom, his hypothesis could not have been tested by his content analysis. Similarly, if the interviews had not revealed anything about people's plans for their bonds, the analysis of the

national cultures through analyzing magazine fiction of each, then the universe might well be all fiction in all magazines appearing in the country during a certain time. Now the sampling problem consists of selecting a representative sample of magazines as well as a representative sample of material from each magazine.

This latter example may serve to illustrate one particularly difficult problem in some kinds of content analysis. If magazine fiction is to be used as a 'reflection' of a nation's culture, should all fiction in all magazines be given equal weight? Should the articles be weighted in some fashion to reflect the number of readers of the magazine? Should articles in a given magazine be weighted according to their placement in the magazine, their length, etc? These questions suggest alternative ways in which the universe of the study might be defined. Any of the procedures might be technically feasible. The choice among them should be governed by the conceptual scheme guiding the research, including in this illustration, such things as the conceptual definition of culture. Is 'culture,' for example, to be defined only in terms of 'symbolic products' or must its description and measurement take into account the number and characteristics of people who are in contact with these products? If he is to be able to justify generalizations from his analyzed content, the investigator must be able to state the rationale for using a given universe of content and to define that universe precisely.

After the universe has been selected for a given investigation, proper procedures for drawing a sample of that universe must be employed. Each unit of the universe must have a known probability of inclusion in the sample, and the *procedure* of selecting units must be independent of any correlations among the units of the universe. These requirements apply both to the selection of sources, if the universe contains more than one source (producer of content, such as respondents in a survey or newspapers, etc), and to the selection of content from any one source.

Let us illustrate some of the dangers that arise because of correlations among units of the universe. Suppose that we have selected as our source a single newspaper and that we are going to sample content from it. Certain procedures that we might employ would produce biased samples, even though we guaranteed that every issue of the newspaper had an equal probability of inclusion in the

to distinguish two rather different types of inference which may be involved in the process. The first type rests upon the assumption that the materials analyzed are a representative sample of some specified universe of (actual or potential) materials. The need to make this kind of inference derives from practical considerations of conducting research. Money and time will be saved if the description of a small sample can be taken as a safe description of the complete universe. The second type of inference rests upon the assumption that the discovered *relations* between certain conditions and certain consequences are universally true. In this type of generalization it is asserted that whenever and wherever the specified conditions obtain, there will follow the specified consequences, and no assumption need be made that the quantitative incidence of certain conditions found in the sample will also be found in the universe.

ASSURING THE REPRESENTATIVENESS OF A SAMPLE In principle a satisfactory system for sampling materials in a content analysis will consist of four elements: (1) specification of the universe to which generalizations are to be made; (2) a guarantee that every unit of the universe has a known probability of inclusion in the sample; (3) a procedure of sampling which is independent of correlations among units of the universe; and (4) a large enough sample to provide a sufficiently small random error of sampling.

The theory and practice of sampling have been extensively developed in recent years, and the reader is referred to Chapter 5 for a systematic discussion of the general problem. Here we shall limit our discussion to some of the more special problems encountered in applying sampling theory to content analysis.

Consider first the problem of specifying the universe of symbolic materials to which generalizations will be made. In any given study the universe that should be selected will depend upon the purposes of the investigation. If, for example, the purpose is to compare the editorial content of a given newspaper against some standard so that legal action may be taken (as in determining the Axis line of native fascist papers during World War II), the universe under consideration should be all editorial content appearing in all issues of that one newspaper over a certain period of time. In this case the problem of sampling is to guarantee that the selected specimens of content accurately represent the total output of the newspaper. If, however, the purpose of the study is to compare

fashion. An illustration of this procedure would be one in which the requirement was set up that one seventh of the sample of newspapers come from each day of the week. Then, if proper methods are employed in sampling the newspapers for each day, one may be sure that each day will be properly represented in the total sample.

ESTABLISHING UNIVERSAL PROPOSITIONS ABOUT RELATIONS BETWEEN CONDITIONS AND CONSEQUENCES The ideal goal of the social psychologist is that he be able to construct universally true statements about relationships among variables. Although his working level of aspiration is ordinarily more modest, his research should nevertheless be designed so that he may approach this ideal. The problems involved in establishing scientific laws are general matters having to do with concept development, hypothesis formation, research design, etc., and cannot be discussed fully in this connection. Our earlier discussions of the problems of determining causation in content analysis and of producing significant findings treat some of the more important considerations especially related to the analysis of qualitative materials. Once a universal proposition has been tentatively formulated, the research task becomes that of replicating the study, seeking limiting conditions, and analyzing apparently exceptional cases.

It may be useful at this point to illustrate the difference between two major types of generalization by means of a specific example. Lazarsfeld, Berelson, and Gaudet (30), in their study of voting behavior in Erie County, Ohio in 1940, found that certain factors such as religion, socioeconomic status, and rural or urban residence predisposed a voter to cast his ballot for one party rather than another. This 'finding' was based upon a sample of interviews in the county. A generalization of this finding to all residents of Erie County rests upon the assumption that the sample employed was representative of the entire county. Since a single county cannot be taken to be a representative sample of all counties in the United States, no safe generalization of this finding can be made to the country as a whole. Furthermore, any generalization to future or past elections with the same county cannot be made safely without further evidence that those conditions producing this predisposition remain constant over time.

In this study another type of 'finding' was also produced. It was discovered that people who were subjected simultaneously to

sample This possibility may be dramatized by an extreme instance Assume that we order the issues as they appeared in time and that we sample one out of every seven issues of the paper Suppose that by some chance procedure we happen to select Sunday as the starting point Now our entire sample will consist of issues appearing on Sunday and we shall have a disproportionately large incidence of features which appear only on Sundays Obviously, this would be a bad sample of the total output of the paper

Many kinds of orderly fluctuations may be found in the content from a given source Mintz (39) has described three major types and has investigated some of the problems of sampling associated with each The first of these he calls primary trends The newspaper treatment of some topic will often show a gradual build up over a period of days If the procedure of sampling happens to select disproportionately from the beginning or ending phases of this trend estimates of the amount of space devoted to the topic will be correspondingly too small or too large It is clear, of course that such trends are not always linear The second type of orderly fluctuation is a cyclical trend An illustration of this kind is the weekly schedule of certain topics in a newspaper or the regular scheduling of certain types of radio programs at certain times of the day If these peaks are either over or undersampled corresponding errors in estimating the total universe will result The third type of orderly fluctuation is where there are compensatory relations between adjoining units Take for example a newspaper in which there is a tendency to give a continuing news story a big play on the first day but little space on the next day If the sampling procedure were to select issues of the paper appearing on alternate days there might result systematic errors in estimating the total universe All these dangers can be minimized by following procedures in which the selection of each sampling unit is independent of the other

In passing it may be noted that under certain conditions there will be an advantage in stratifying the universe That is to say whenever there is reason to believe that certain classes of units may be more homogeneous than the total universe these classes may be designated and the requirement set up that the sample contain a proper proportion of units from each class In these circumstances the selection of each unit should of course still be done in a random

affairs and attitude toward the United Nations, the analyst must specify before he constructs his outline just what data he will take to test this relation. He might decide that he will want to present in his report a matrix in which the columns indicate several positions on an attitude scale and the rows show different scores on an information test. He may want to present the frequency of interviews falling in each cell and test whether the distribution differs significantly from a random one. A similar specification of needed data should be made for the entire investigation.

Step 2 Map Out Plans for Tabulation

A great deal of trouble can be avoided by making explicit plans for the tabulation of coded data before constructing the analysis outline. It makes a good deal of difference, for example, whether the coded data are to be punched on cards for machine processing or to be tabulated by hand. Although the variables and the categories of the outline will not usually be different for different methods of tabulation, their arrangement within the outline and the system of notation employed in coding may well be quite different. Since tabulation by punch cards is accomplished by punching a numbered position in a numbered column, the appropriate notations on coding sheets consist of indicating a variable by a column number (or numbers) and a category by a number within the column. Thus, 'attitude toward the United Nations' might be assigned to column "27." In this column a favorable attitude might be given the number '1,' a neutral attitude number '2,' and an unfavorable attitude number '3.' When there is insufficient evidence for making any rating, the number '0' might be noted in the same column. If tabulation is to be done by machine, it will be helpful to consult, at the time the analysis outline is being constructed, a person experienced in machine tabulations so that the many 'tricks of the trade' and short cuts which are possible can be built into the analysis outline.

Step 3 Lay Out the Skeleton of the Outline

At this point, it will be useful to list the variables in terms of which the content is to be coded. If the investigation consists in

conflicting predisposing factors (such as a rural Catholic or a low income Protestant) displayed various symptoms of conflict in making a political decision. For example, they took longer to make up their minds and showed more vacillation in their party preference. The theory that conflicting forces or cross pressures produce such symptoms of conflict may be proposed as a universal "law" which should hold wherever or however such cross pressures are exerted. The truth or falsity of this theory does not depend upon the representativeness of the sample employed in the study, any valid exception to it found anywhere would be sufficient to require a modification of the proposition.

CONSTRUCTING THE ANALYSIS OUTLINE

The preceding discussion has examined the major systematic principles involved in converting phenomena into scientific data. In conducting research, more is needed, however, than an understanding of these fundamental principles. The success of any project will depend upon the degree to which these principles are expressed as actual procedures. Let us turn then, to a consideration of some of the more concrete and detailed procedures involved in carrying out investigations that employ content analysis.

How, specifically, does one go about constructing an analysis outline? Six steps for arriving at a satisfactory analysis outline may be indicated. These steps are intended to suggest clusters of interrelated decisions which the analyst must make. They are points at which it is useful to check the emerging outline against the general principles listed above.

Step 1 Specify Needed Data

In laying out an analysis outline, it is essential that the investigator have clearly in mind specifically what data are required by his total research design. Ordinarily, he will encounter less difficulty in the long run if he is able at this point to work out his plans in sufficient detail so that he can tell what form his final tables will take. If, for example, his design calls for testing in a set of interviews the relationships between information about international

- 4 Device—what is the rhetorical or propagandistic character of the communication?

Step 4 Fill in Categories for Each Variable

There are many systems of categories which may be employed for any given variable. The one chosen will depend upon the objectives of the study and the type of measurement being undertaken. Whatever type of system is chosen, the analyst should check to see that it meets what Lazarsfeld and Barton (29) call "the requirement of logical correctness." A system of categories meets this requirement if it is exhaustive and if its categories are mutually exclusive. It is exhaustive if there is a category in which to place every relevant item which may be found in the content. Its categories are mutually exclusive if there is one and only one place to put an item within that system of categories. Although this requirement of logical correctness appears simple and obvious, it is remarkable how frequently it is violated. Experience indicates that it will be well to check each system of categories before they are finally used, to be sure that it is satisfactory in this respect. Systems of categories which call for listings of themes, reasons, arguments, sources of influence, and the like seem especially vulnerable to this type of error. The following classification of places where people were solicited to buy war bonds is not a far fetched example. place of work, home, store, bank, post office. Now, this system of categories is neither exhaustive nor mutually exclusive. Obviously, there are other places where solicitation might take place—and where would one categorize a farmer who was solicited at home on his farm?

In constructing categories, one is often confronted with a dilemma. If a category is too broad, it conveys little specific meaning, but if it is too narrow, the coded material differs little from the "raw" material. One resolution of this dilemma is through use of grouped categories. Thus, a system of categories for classifying reasons for buying war bonds might designate such broad categories as "Personal financial," "National patriotic," "National economic," etc. Then, under each heading there might be more specific categories, such as "Bonds are safe investment," "Bonds pay good rate of interest," "Money invested in bonds is exempt from temptations of spending," etc. In the interpretation of findings from the

analyzing interviews, these variables will be used to classify not only various features of the answers to questions about the respondent's psychological make-up but also such matters as his age, income, marital status, and other demographic or behavioral characteristics. In listing the variables to be included in the outline, care should be taken to assure that *all* information needed on the punch cards is placed on some variable. Thus, the outline should contain provision for coding the name of the study, the number of each enumeration unit (interview, issue of a newspaper, etc.), the name of each coder, and any information relevant to tests of reliability or other statistical treatment.

In the literature on content analysis of communication materials, certain types of variables have been employed rather frequently. These have been summarized by Berelson (6) under the two broad headings of "What is said" and "How it is said." The variables listed by him under each are given here, in order to indicate some of the kinds of variables that one might profitably employ.

A WHAT IS SAID

- 1 Subject matter—what is the communication about?
- 2 Direction—is the treatment favorable or unfavorable toward the subject?
- 3 Standard—what is the basis (or grounds) on which the classification of direction is made?
- 4 Values—what goals are explicitly or implicitly revealed?
- 5 Methods—what means or actions are employed to realize goals?
- 6 Traits—what characteristics of persons are revealed?
- 7 Actor—who initiates actions?
- 8 Authority—in whose name are statements made?
- 9 Origin—what is the place of origin of the communication?
- 10 Target—to whom is the communication particularly directed?

B HOW IT IS SAID

- 1 Form of communication—is it fiction, news, television, etc.?
- 2 Form of statement—what is the grammatical or syntactical form of the unit of analysis?
- 3 Intensity—how much strength or excitement value does the communication have?

single assertion (3) the smallest segment of content required to yield a single characterization, such as an adjectival phrase, value judgment, and the like, (4) a character, person, group, or institution that is described in the content, (5) a paragraph or other natural unit of meaning, and (6) an item such as an article speech, radio program, etc

In analyzing free answer interviews, the most frequently used recording unit is the answer to a single question. It is not uncommon, however, to use larger or smaller units. For certain purposes an entire interview may be taken as a single unit and characterized as a whole. For other purposes a certain set of questions may be treated as one unit. Or there may be good reason to break down the answer to a single question into units consisting of single words themes value judgments, or reasons.

Designation of the context unit is often left quite vague or to the individual coder's judgment. Since the major purpose in setting up a context unit larger than the recording unit is to provide better bases for perceiving the 'meaning' of the recording unit, there seems to be some justification in allowing the coder to seek clarification throughout the material. Such a procedure, however, sometimes greatly reduces the reliability of coding. Whenever it is possible, the coder should be given quite specific instructions somewhat like the following: Read the answers to questions 2, 3, and 4 before categorizing the reasons given in question 5, but do not read the answers to questions coming later in the interview, or 'Read an entire paragraph, but no more, before coding the value judgments within the paragraph.'

The unit of enumeration that seems to be most popular in communication research is that of physical length (such as column inch etc) or temporal duration. If such units are meaningful from a theoretical point of view, they should be used, because they have real advantages of reliability and susceptibility to mathematical manipulations. In interview surveys, the most commonly used unit of enumeration is the respondent. This, too, is a convenient unit because it is ordinarily safe to consider each respondent as quantitatively equal to every other. If there is some theoretical reason for not treating each respondent equally, other units of enumeration may be required. For example, because of the functional interdependence of several people who are supported financially by the

study, the investigator may utilize each level of classification for different purposes

If the analysis outline contains a considerable number of variables, it is likely that rather similar systems of categories will be found among these variables. In the interest of coding speed and the reduction of errors, it has been found desirable to establish certain consistencies in the way in which the categories are arranged. The University of Michigan Survey Research Center (44) has established certain conventions which it follows in categorizing free answer interviews. For example, the category 'yes' is always given the code number "1," the category 'no' number "5," "don't know" number 9, none number '0,' etc. In a similar way, it has been found desirable to standardize the numbering system for scales so that they all progress in the same direction from positive to negative or from high to low. With such standardization, the coder can soon categorize material almost automatically.

When the analysis outline has been completed, with all the categories defined, a manual of instructions for coders should be written giving these definitions in clear operational terms.

Step 5 Establish Procedure for Utilizing the Material

We have defined above three kinds of units which must be dealt with in any content analysis: the recording unit, the context unit, and the unit of enumeration. The specific working definitions to be used in the study should be established at this point in such a fashion that various coders can all unitize the same material in the same way. These definitions should be written down as a part of the coding instructions. The selection of definitions of these units should be guided by the same theoretical framework that determines the rest of the research design. "Practical" considerations of coding efficiency and reliability should not be ignored in deciding such things as the "size" of unit, but valid coding depends upon the theoretically correct selection of units whose categorization can properly be taken to indicate some significant feature of the material.

The most common recording units in communication research are (1) a single word, (2) a theme, usually consisting of a subject and predicate or some larger unit which can be condensed into a

USING THE ANALYSIS OUTLINE

If the content analyst has skillfully taken the steps indicated above, he should now have an analysis outline well suited to his research objectives, appropriate to the content at hand, and amenable to efficient tabulation and statistical treatment. The remaining requirement is that he have coders who are able to use the analysis outline as intended and in a standardized manner.³ It is useful to think of the coder as a measuring instrument which must be sensitive to variations in the material and dependable in the sense that it responds in the same way to functionally equivalent content. In order to have coders who possess these characteristics, it is necessary to select people with the proper abilities, to train them adequately, and to supervise their work effectively.

Selection of Coders

For satisfactory coding, certain skills and abilities are essential. The coder must be a sensitive person, well differentiated in respect to symbolic materials. He must be able to detect subtle differences of meaning but also to neglect differences that do not make a difference for a specific purpose. In other words, he must be able to make use of the genotypic categories required by the analysis outline. In most social-psychological research, this means that the coder must have some acquaintance with the concepts of social psychology. If the analysis outline requires only phenotypic categories or categories defined in terms of everyday usage, the coder may well be an intelligent layman. A reasonably good level of intelligence is the minimal requirement for any content analysis.

If the quantity of material to be coded is great, an additional requirement must be met. The process of coding involves the repetitive application of the analysis outline to the material. Reliable

³ This discussion of how to use the analysis outline is written on the assumption that several people will do the actual coding. If the volume of material to be analyzed is large, this assumption is realistic. Even, however, when the analyst and the coder are the same person, it is desirable to have an independent coder so that the whole procedure can be objectified. Unless the analyst forces himself to communicate his definitions of categories, units, etc. to an independent coder, he can have little assurance that his procedures are in fact reproducible.

same source, it is desirable in some kinds of economic surveys to employ a "spending unit" as the unit of enumeration. Information may be obtained separately from each individual, but it would be pooled into one "spending unit" in order to construct a single unit for purposes of computing frequencies, means, and distributions.

Step 6. Try Out the Analysis Outline and Unitizing Procedure

After the analysis outline and the unitizing procedure have been developed, they must be applied to the content in a preliminary way in order to discover what modifications are needed. Ordinarily, this trying out of the coding procedures is also used as a training period for those people who are to do the final coding. When this period is ended, the analysis outline should be fixed in its final form and the coders should be "set" in their use of the coding procedures.

This stage has been standardized at the University of Michigan Survey Research Center (44) in a procedure known as the "Round Robin." A random set of materials is collected, and each coder codes it independently. All disagreements among coders are noted and used as a preliminary check on the reliability of coding. These disagreements are also examined to see what improvements in the analysis outline should be made. It is not uncommon to make substantial modifications at this point. Variables of the analysis outline which do not fit the material well will need to be redefined or eliminated. Systems of categories which are either not exhaustive or not mutually exclusive will be detected and revised if the Round Robin is done well. Those variables made up of listings can be expanded so that few additions to the list will be made after "production" coding gets under way. And, finally, the system of notation on the coding sheets will be checked to determine whether it is most convenient for rapid and automatic coding and whether it will facilitate tabulation as much as possible.

When the Round Robin is completed, the whole coding procedure should be frozen so that the entire set of content will be coded in the same way. Any modifications of the analysis outline after final coding has begun must be made retroactive to all materials coded prior to the change. Obviously, much time could be spent in making such changes if they were to occur very often.

periodic coding, and when the task can provide satisfaction other than merely economic gain. College students who combine the financial incentive with a larger goal of training or social service and who do coding as a part time occupation appear ideally suited to this kind of task.

Training of Coders

Once the coders have been assembled for a given project, it is necessary to train them in the use of the analysis outline. In general, it is desirable to communicate to the coders a full understanding of the purposes of the project—why it is being done, what uses will be made of the findings, and any other motivations which activate the project director in undertaking the study. A full understanding of these matters on the part of the coders will allow them to do their work more intelligently and with a higher level of motivation. Unless these matters are communicated effectively, many decisions made by the project director will appear meaningless and arbitrary to the coders. There may be, upon occasion, specific hypotheses which should be kept from the coders for fear that such information might 'contaminate' the coding, but decisions to withhold information from the coders should always be made only when other procedures cannot be followed equally well. Coders who better understand how the analysis outline was constructed will be better able to adopt the rationale behind the operational definitions of categories and units.

After the general purposes of the project have been communicated, instruction in the details of the outline may begin. The purpose of this training is to establish a common frame of reference and common operational definitions among all the coders. It is well to begin this phase of training with oral and written descriptions of the variables and categories. Then after the formal definitions have been grasped, the coders should begin trying out these definitions on the materials. At this stage it is desirable to move into the Round Robin. As mentioned above, the independent coding of the same materials by all the coders serves the dual purpose of revising the analysis outline and of standardizing the coders. It is essential that the Round Robin be conducted with enough different types of content and with sufficient discussion so that all major problems that

coding demands, therefore, that the outline be used in the same way (the same operational definition of categories, the same frame of reference, the same degree of differentiation, the same level of attention to details, etc.) throughout the entire coding operation. A person who is easily satiated with repetitive work will consequently not make a good coder for a very long period of time. Studies of satiation by Karsten (24), Kounin (25), and others have shown that the very requirements of sensitivity, motivation, and deep involvement in the task tend to hasten satiation if the meaning of the task is that of sheer repetition. Such satiation produces errors and variability in the application of the analysis outline to the materials. Unfortunately, since no good test of susceptibility to satiation has been developed, there is little that can be done at present to minimize this problem through differential selection of coders. It would appear, however, that people who view the task of coding as routine, menial work will view it as mere repetition of the "same" activity and will, as a result, be satiated more readily.

If the coding is to be carried out by a team of coders, it is necessary that they all come to apply the same definitions and frame of reference to the coding. The achievement of such a common approach to the task is best accomplished through group discussions, the Round Robin, and the sharing of difficult coding decisions. A coder who is uncommunicative or ego-defensive will, therefore, not contribute well to this purpose and will probably heighten the unreliability of coding. Again, it must be pointed out that no very satisfactory objective test of these personality traits now exists, and that selection for these traits is difficult.

In present practice it appears that good coders are discovered mainly through a process of selection "under fire," and that some provision might well be made to begin with a somewhat larger staff of coders than will "survive" to the end of the project.

Large research organizations who maintain a permanent coding staff have found it difficult to maintain the same people over a period of years at a high level of morale. Sensitive and intelligent people who are acquainted with the concepts of social science rarely find it satisfying to make a life career of such repetitive and routine work. Rarely can such a person work full-time at such a job for more than a year or two without considerable demoralization. Much better morale seems to result when the arrangement is for part-time or

In using catch-all categories, it is sometimes useful to keep a separate hand tabulation of the different items that are thus coded. For example, suppose that a system of categories contains the category "other reasons." Each time a recording unit is placed in the category, a notation is made of the nature of the specific reason, along with an identification of the unit. Then, if it is discovered that some specific reason is appearing with a considerable frequency, it may be separated from the "other reasons" category and tabulated by itself.

As coding proceeds, it is important to hold periodic discussions among the coders to assure that the same frame of reference and operational definitions of categories are maintained throughout the coding period. The entire group should discuss any persisting disagreements in the use of certain categories, and other problems arising out of experience with the analysis outline.

The reliability of coding can be measured and stability of coding promoted by use of "check coding." In this procedure a certain percentage of the content is independently recoded by a "check coder." The check coder may be someone who is taken as a sort of criterion (perhaps the principal investigator), or a system similar to the Round Robin may be used in which each coder serves as a check coder for each of the others. After the check coder has finished a set of material previously coded by one coder, the two should get together to discuss each of their disagreements. If the discussion carries an atmosphere conducive to learning rather than self-justification, this discussion can serve well to improve the quality of coding as the study goes along.

Records of disagreements turned up in the check coding should be kept and tabulated in various ways. These records, of course, can be used as one type of measure of the total reliability of coding. They can also, however, be broken down into several more specific analyses. The various variables of the analysis outline should be examined separately to determine whether some of them are producing abnormally high unreliability. Each of these variables may also be examined more minutely to determine precisely what kinds of disagreements were most common within a single variable. It may be found, for example, that coders simply could not distinguish reliably between two adjacent categories and that the two categories should be merged into one for further tabulation. Finally, the

might subsequently arise are worked out. A running record of coder disagreements is essential as an aid in revising the outline and as an indication of when the Round Robin may be safely terminated. When, by actual performance, the coders have demonstrated their ability to code reliably in the way called for by the research design, 'real' coding may finally begin.

Mechanics of Coding

The orderly processing of materials requires that regularized procedures be established for their storage, assignment to coders, and recording. It has been found convenient to package together a collection of materials (perhaps ten interviews or ten issues of a newspaper) and to have a coder 'check out' a package at a time. When one package has been coded, it is returned to a central storage place and a new package is taken. In the assignment of materials to coders, it is desirable to randomize the materials so that any systematic biases among coders will not appear as trends or correlations in the coded data. An orderly assignment of materials to coders will also assure, of course, that all materials get coded once and only once. Similar care should be exercised in collecting and storing the sheets upon which coding has been recorded.

Even after the Round Robin has been completed and final coding has begun, it may be found necessary to add new categories to some of the variables of the outline. Or it may be discovered that too many items are falling into such catch-all categories as 'other reasons'. If a coder comes across a recording unit for which there is no category, he should bring this case to the attention of the coding supervisor. The supervisor should assess the merit of adding a new category by determining whether the new category would be meaningful within the rationale of the system of categories involved by ascertaining whether the case in point could not just as well be placed in an existing category and by judging whether the new category would be used frequently enough to warrant its separate designation outside one of the catch-all categories. If the decision is made to modify the analysis outline, this change must be made on the outline being used by all coders. Furthermore, some procedure must be established to guarantee that the coding of all previously coded materials is modified wherever required.

for the social scientist. It should be viewed, however, only as a tool. Even when it is extremely well fashioned, its scientific or practical value may, in any specific project, turn out to be negligible. A successful research project will combine both technical excellence and a good research design aimed at answering significant research questions.

BIBLIOGRAPHY

- 1 Allport, G. W. *The use of personal documents in psychological science*. New York: Social Science Research Council, 1942.
- 2 Arnheim, R. The world of the daytime serial. In Lazarsfeld, P. F. and Stanton, F. (eds.) *Radio Research 1942-43*. New York: Duell Sloan and Pearce, 1944, pp. 34-107.
- 3 Baldwin, A. L. Personal structure analysis: a statistical method for investigating the single personality. *J. Abnorm. Soc. Psychol.*, 1942, 37, 163-183.
- 4 Bales, R. F. *Interaction process analysis: a method for the study of small groups*. Cambridge: Addison-Wesley Press, 1950.
- 5 Berelson, B. The effects of print upon public opinion. In Waples, D. (ed.) *Print, radio, and film in a democracy*. Chicago: Univ. of Chicago Press, 1942, pp. 41-65.
- 6 ———. *Content analysis in communication research*. Glencoe: The Free Press, 1952.
- 7 ———, and Salter, P. Majority and minority Americans: an analysis of magazine fiction. *Publ. Opin. Quart.*, 1946, 10, 168-190.
- 8 Britt, S. H., and Lowry, R. L. Conformity behavior of labor news papers with respect to the AFL-CIO conflict. *J. Soc. Psychol.*, 1941, 14, 375-387.
- 9 Bruner, J. S., and Allport, G. W. Fifty years of change in American psychology. *Psychol. Bull.*, 1940, 37, 757-776.
- 10 Bureau of Agricultural Economics. *Men's preferences among wool suits, coats, and jackets*. Agricultural Information Bulletin No. 64. Washington: United States Department of Agriculture, 1951.
- 11 Cantril, H. Public opinion in flux. *The Annals*, 1942, 220, 136-152.

records should be tabulated separately for each coder in order to yield a measure of his reliability as a coding instrument. The interpretation of differences among coders should be made judiciously, of course, because it is possible that the coder with the greatest number of disagreements might be the most 'valid' coder.

The proportion of the material which should be check coded will depend upon the uses that are to be made of it. For purposes of training and of maintenance of constant standards, there is some point in check coding a relatively larger proportion earlier in the coding process and to taper off as coding becomes stabilized. In order to construct a measure of reliability, it is best to employ a random sample of all materials. In some cases, where any error is deemed serious, it may be desirable to check code the entire set of material and to tabulate only the pooled judgment.

SUMMARY

The fundamental objective of all content analysis is to convert phenomena (*i.e.*, symbolic behavior of people) into scientific data. We have specified four characteristics which scientific data must display: (1) objectivity and reproducibility, (2) susceptibility to measurement and quantification, (3) significance for systematic theory, either 'pure' or 'applied', and (4) generalizability.

In constructing an analysis outline for a given project, it will be useful to organize the work so that it consists of six steps, or clusters of decisions. At each of these points the developing outline should be checked against the formal requirements for scientific data. These steps are: (1) specifying needed data, (2) mapping out plans for tabulation, (3) laying out the skeleton of the outline, (4) filling in categories for each variable, (5) establishing procedure for unitizing the material, (6) trying out the analysis outline and unitizing procedure on a sample of the material.

The successful use of a well developed outline depends upon the selection of capable coders, effective training of them in the outline being used, and the establishment of good supervision so that proper procedures of coding are followed.

Experience over a number of years with content analysis reveals that when technically well executed it can be a most valuable tool.

- 29 Lazarsfeld P and Barton A H Qualitative measurement in the social sciences classification typologies and indices In Lerner D and Lasswell H D (eds) *The policy sciences* Stanford Stanford Univ Press 1951 pp 155 192
- 30 Lazarsfeld P F Berelson B and Gaudet H *The people's choice how the voter makes up his mind in a presidential campaign* New York Duell Sloan and Pearce 1944
- 31 Lee A M and Lee E B (eds) *The fine art of propaganda a study of Father Coughlin's speeches* New York Harcourt 1939
- 32 Leites N Bernaut E and Garthoff R L Politburo images of Stalin *World Politics* 1951 3 317 339
- 33 Lerner D The American soldier and the public In Merton R and Lazarsfeld P (eds) *Continuities in social research* Glencoe Free Press 1950 pp 212 247
- 34 Lewin H S Hitler Youth and the Boy Scouts of America a comparison of aims *Hum Relat* 1947 1 206 227
- 35 McGranahan D V and Wayne I German and American traits reflected in popular drama *Hum Relat* 1948 1 429 455
- 36 Merton R K *Mass persuasion the social psychology of a war bond drive* New York Harper 1946
- 37 Miles J *The continuity of poetic language studies in English poetry from the 1540's to the 1940's* Berkeley Univ of Calif Press 1951
- 38 Millsbaugh M Trial by mass media? *Publ Opin Quart* 1949 13 328 329
- 39 Mintz A The feasibility of the use of samples in content analysis In Lasswell H D and Leites N (eds) *Language of politics* New York Stewart 1949 pp 127 152
- 40 Ojemann R H et al A functional analysis of child development material in current newspapers and magazines *Child Devel* 1948 19 76 92
- 41 Royal Commission on the Press 1947 1949 *Report* London His Majesty's Stationer's Office 1949
- 42 Sanford F H and Rosenstock I M Projective techniques on the doorstep *J Abnorm Soc Psychol* 1952 47 3 16
- 43 Skinner B F The alliteration in Shakespeare's sonnets a study in literary behavior *Psychol Rec* 1939 3 186 192
- 44 Survey Research Center University of Michigan *A manual for coders* Ann Arbor Institute for Social Research 1952

- 2 Cartwright D Relation of decision time to the categories of response
Amer J Psychol 1941 54, 174 196
- 13 ——— Some principles of mass persuasion *Hum Relat*, 1949 2,
253 267
- 14 ——— Survey research psychological economics In Miller J G
(ed) *Experiments in social process* New York McGraw Hill 1950
pp 47 64
- 15 ——— and Festinger L A quantitative theory of decision *Psychol
Rev* 1943 50 595 621
- 16 Crutchfield R S and Gordon D A Variations in respondents inter-
pretations of an opinion poll question *Int J Opin and Attitude
Res* 1947 1 No 3 1 12
- 17 Dollard J and Mowrer C H A method of measuring tension in
written documents *J Abnorm Soc Psychol* 1947 42 1 32
- 18 Festinger L Cartwright D Barber K Fleischl J Gottsdanker J
Keyser A and Leavitt G A study of rumor its origin and spread
Hum Relat 1948 1 464 486
- 19 Flesch R *How to test readability* New York Harper 1951
- 20 George A *The intelligence value of content analysis* Unpublished
ms 1951
- 21 Hart A *Shakespeare and the homolies* Melbourne Melbourne Univ
Press 1934
- 22 Hart H Changing social attitudes and interests In Report of the
President's Research Committee on Social Trends *Recent social trends
in the United States* Vol 1 pp 392 442 New York McGraw Hill
1933
- 23 Janis I R and Fadner R H The coefficient of imbalance In Lass-
well H D and Leites N (eds) *Language of politics* New York
Stewart 1949 pp 153 169
- 24 Karsten A Psychische Sättigung *Psychol Jorsch* 1928 10 142 204
- 25 Kounin J S Experimental studies of rigidity *Character and Pers*
1941 9 251 282
- 26 Lasswell H D *Propaganda technique in the world war* New York
Knopf 1927
- 27 ——— Detection propaganda detection and the courts In Lasswell
H and Leites N (eds) *Language of politics* New York Stewart
1949 pp 173 232
- 28 ——— and Blumenstock D *World revolutionary propaganda* New
York Knopf 1939

Theory and Methods of Social Measurement

*Clyde H. Coombs*¹

What one "finds out" from one's data is a function of two things: the information in the data and how this information is extracted. What information the data contain depends on how it is collected. Some methods of collecting data "permit" more characteristics of behavior to exhibit themselves than do other methods. Or, in opposite terms, some methods of collecting data impose properties on the behavior that other methods do not. Obviously, properties imposed on the data by the method of observation cannot be inferred to be properties of the behavior in question.

The method of analysis, then, *defines* what the information is and may or may not endow this information with certain properties. A "strong" method of analysis endows the data with properties which permit the information in the data to be used, for example, to construct a unidimensional scale. Obviously, again, such a scale cannot be inferred to be a characteristic of the behavior in question if it is a *necessary* consequence of the method of analysis.

It therefore becomes desirable to study methods of collecting

¹ I wish to thank Leon Festinger, Howard Raiffa, and Robert Thrall for reading the manuscript of this chapter and contributing many valuable criticisms and suggestions.

- 45 Sussmann L. A Labor in the radio news an analysis of content
Journalism Quart 1945 22 207 214
- 46 US Strategic Bombing Survey *The effects of strategic bombing on
German morale II* 1946 US Gov t Printing Office
- 47 Waples D and Berelson B *What the voters were told An essay in
content analysis* Univ of Chicago Graduate Library School 1941
Mimeographed
- 48 White R K Hitler Roosevelt and the nature of war propaganda
J Abnorm Soc Psychol 1949 44 157 174
- 49 ——— *Value analysis the nature and use of the method* New York
Society for the Psychological Study of Social Issues 1951
- 50 Wolfenstein M and Leites N *Movies a psychological study* Glencoe
Free Press 1950
- 51 Woodward J L *Foreign news in American morning newspapers a
study in public opinion* New York Columbia Univ Press 1930
- 52 Yakobson S and Lasswell H D Trend May Day slogans in Soviet
Russia 1918 1943 In Lasswell H D and Leites N (eds) *Language
of Politics* New York Stewart 1949 pp 233 297
- 53 Yule G U *The statistical study of literary vocabulary* Cambridge
England The University Press 1944

Theory and Methods of Social Measurement

*Clyde H. Coombs*¹

What one "finds out" from one's data is a function of two things: the information in the data and how this information is extracted. What information the data contain depends on how it is collected. Some methods of collecting data "permit" more characteristics of behavior to exhibit themselves than do other methods. Or, in opposite terms, some methods of collecting data impose properties on the behavior that other methods do not. Obviously, properties imposed on the data by the method of observation cannot be inferred to be properties of the behavior in question.

The method of analysis, then, *defines* what the information is and may or may not endow this information with certain properties. A "strong" method of analysis endows the data with properties which permit the information in the data to be used, for example, to construct a unidimensional scale. Obviously, again, such a scale cannot be inferred to be a characteristic of the behavior in question if it is a *necessary* consequence of the method of analysis.

It therefore becomes desirable to study methods of collecting

¹ I wish to thank Leon Festinger, Howard Raiffa, and Robert Thrall for reading the manuscript of this chapter and contributing many valuable criticisms and suggestions.

data with respect to the amount and kind of information each method *contains* about the behavior in question as distinct from that *imposed*. Similarly, it becomes desirable to study the various methods of analyzing data in terms of the characteristics or properties each method imposes on the information in the data as a necessary preliminary to extracting it.

All of this is a part of measurement theory, a subject of greater concern in the social sciences than it is in many other domains of knowledge. Measurement in the physical sciences usually means the assigning of numbers to observations (a process called 'mapping'), and the analysis of the data consists in manipulating or operating on these numbers. The social scientist taking physics as his model, has, frequently, attempted to do the same. It is the thesis of this chapter that the social scientist who follows such a procedure will sometimes violate his data.

WHAT IS MEANT BY MEASUREMENT

The objective of this first section is to show that the theory of measurement consists of a system of distinct theories, each corresponding to what may be called a *level* of measurement, and that a given set of data may satisfy (permit the valid use of) some of these levels of measurement but not others. This first section will be concerned with an incomplete generalization of the logic of measurement, with examples at each of the levels discussed. This will be followed by a section on a theory of data in which the distinction between collecting and analyzing data will be discussed in terms of certain abstracted invariants of the behavior of individuals. This theory of data is an effort to construct a framework in terms of which all methods of collecting and analyzing data may be unified under a general system. In the final two sections, methods of collecting and analyzing social psychological data will be discussed in the context of this general theory of measurement and the theory of data.

Throughout this chapter, the major emphasis will *not* be on the application of specific techniques. This type of material is available throughout the literature. Instead, emphasis will be placed on the assumptions underlying the several techniques and their structure.

and interrelations. An understanding of these aspects of the nature of measurement is necessary for an intelligent choice of a method of collecting data and of a method of analysis. The various methods of collecting data contain information which may differ both quantitatively and qualitatively. Similarly, the various methods of analyzing data may differ in the degree to which they rely on information contained in the data as contrasted with the structure and relationships they impose on the data.

In the discussion to follow, some of the various levels of measurement within the general theory of measurement will be indicated. The nominal scale, which is the simplest possible level of measurement, will be discussed first.²

The Nominal Scale

Measurement in its simplest form consists of substituting symbols or names for real objects. When measurement consists only in this mapping of objects into symbols, the symbols constitute a nominal scale (29). Thus a system of occupational families or psychiatric classifications is an attempt to construct a nominal scale. A nominal scale has certain properties which may be formulated abstractly as axioms. For example, either the relation of "equal to" or "not equal to" must hold between objects on a nominal scale. This means that any pair of objects must clearly belong to the same class or not belong to the same class. In addition, the relation of equality must be symmetric and transitive. By symmetry is meant that if the relation holds between a and b , it also holds between b and a , symbolically, if $a = b$, then $b = a$. By transitivity is meant that if $a = b$ and $b = c$, then $a = c$.

This level of measurement is so primitive that it is not always recognized as measurement, but it is a necessary condition for all higher levels of measurement.

The psychological processes of perception are representative of measurement on a nominal scale. Perception may be regarded as the mapping of stimuli into equivalence classes. The properties of the "ideal" of an equivalence class then become the properties of a specific object, and such phenomena as size constancy and a great

² A more detailed generalization of the logic of measurement theory is contained in Coombs (8 Chap. 1).


variety of other constancies are to be anticipated. The mappings of individuals into such classes as athletes, politicians, Negroes, etc., are examples of stereotyping and constitute nominal scales. Cultural and educational factors affect the construction of such nominal classes, creating new ones, dismembering old ones, and creating relationships among classes. The relationship between nominal scales and perception is so close that whether or not an individual perceives and what he perceives is dependent in the first place upon the existence of nominal classes and in the second place upon the range or spread of the classes.

The Partially Ordered Scale

Sometimes the objects in one class of a nominal scale are more than just *different from* the members of another class—they may bear some kind of a *relationship to* them. One such relationship is that the members of the one class are *more* of something than the members of the other class and it is meaningful to say that the relation greater than ($>$), in some respect, holds between the members of one class and the members of the other. Given a number of equivalence classes, if such a relation holds between *some* pairs of classes the result is a partially ordered scale. For example, suppose one wants to measure something to be called socioeconomic status. Let us also suppose, for the sake of simplicity of illustration, that this attribute is made up of income and educational level. If individual A has more income and *at the same time* more education than individual B , he can be said, then, to have a higher socioeconomic status than B , ($A > B$). Further, if B has more of both constituent attributes than a third individual C , then not only is $B > C$, but also A has higher socioeconomic status than C , ($A > C$). It is apparent, then, that this relation is transitive. It is asymmetric, however, because if $A > B$, then $B \not> A$.

Suppose now that there were some fourth individual, D , who had less of both attributes than the first individual, A , and more of both attributes than the third individual C . It could also be said then, that, with respect to socioeconomic status, $A > D > C$. But suppose at the same time that, although D has more income than B , he has less education. This poses a problem. It is not immediately clear whether $B > D$ or $D > B$ with respect to socioeconomic status.

or, in fact, whether either of these relationships exists. If it is not insisted that one of the relationships above must necessarily hold, then this pair of individuals, B and D , may be spoken of as being incomparable, and the scale of socioeconomic status of A , B , C , and

D constitutes a partial order of the form , where the connecting line between two individuals signifies that the higher individual has more status than the lower individual. Where there is no connecting line the two individuals are not comparable. Hence in this partial order it cannot be said that B has more status than D or vice versa. This level of measurement is called a partially ordered scale and may be obtained whenever an attribute is made up of two or more primitive attributes which do not combine additively or compensate for each other.

The Ordinal Scale

Beginning with a nominal scale and finding a relation (e.g., $>$) to hold for some pairs, one can construct a partially ordered scale, a higher level of measurement. If one finds that the relationship holds for *all* pairs of objects from different classes, one can make an appropriate change in the axioms to convert a partially ordered scale into a simply ordered scale or what is usually known as the ordinal scale.

Let us return to the question of socioeconomic status and see what is required to construct an ordinal scale in that case. We already have an ordinal scale there in all respects except that individuals B and D are not comparable. To construct an ordinal scale, all pairs must be comparable, so obviously individuals B and D must be 'found' to be comparable or made to be comparable. This can be done by equating one of the constituent attributes to the other. An example would be to equate \$1000 of income to one year of education. This would be a simple linear transformation but any system would do. In the absence of any natural basis a decision is made concerning these equivalences which relates each unit of one to every unit of the other. Immediately, then, there would be a simple operational basis for comparing all pairs of individuals or for placing them in equivalence classes. An ordinal scale has thus been constructed.

Suppose one were interested in measuring the popularity of individuals and one observed 'who likes whom'. The relationship $A > B$ means that individual A is liked by the same people who liked individual B and some others in addition. The data might be expected to yield a partial order. To construct a simple order, one needs only to assume that being liked by one individual is equivalent to being liked by any other individual. Then immediately the popularity of A is measured by the number of people who like A and at least an ordinal scale has been constructed (a certain abstract equivalence to the procedures of mental testing should be evident here).

The process indicated here is not an unusual one. It is because of precisely such procedures that operationism (2) was given so much attention. Operationism is the principle that the meaning to be given to a concept resides in the operations which give rise to the measure of the concept. Aside from the philosophical reservations one might have about operationism, the warning must be given in passing to avoid operationism *in reverse*. By operationism in reverse is meant endowing the measures with all the meanings associated with the concept. In the preceding example the concept of socioeconomic status would mean income, with a unit of education counted as so many dollars. A different operational definition would be a different concept—the equivalence of education to dollars is part of the definition.

Attempts were made above to construct an ordinal scale by mapping a partially ordered attribute into a simple order. We saw that this mapping required a decision to be made on the exchange rate between the components and that every different exchange rate would result in a different simple order for a large population. If the experimenter insists on mapping a partial order into a simple order, the best he can usually do is to set up an objective communicable rule for the substituting of components and impose it on the data.

One alternative procedure is to study the partial order itself. It is possible, for example, to build a partial order by combining two or more simple orders in certain ways. Thus there is the suggestion that a partial order may be decomposed into simple orders, a process analogous to factor analysis. This problem appears soluble (10) and constitutes a generalization of multiple factor analysis.

to nonmetric systems. The determination of the dimensionality of a partial order (corresponding in principle to the rank of a correlation matrix) already appears possible, but routine procedures for decomposing a partial order into alternative sets of simple orders (corresponding in principle to the rotational problem of factor analysis) have still to be developed.

A "natural" ordinal scale is obtained when the raw data themselves contain the relation "greater than" for all pairs. It is remarkably difficult to find an example of a simple order among social psychological variables, since partial orders are almost universal. Let us suppose, however, that we wish to measure the "authority" of some military personnel of various ranks and that instead of deriving the scale from military regulations we shall derive it from observational data. Let us take "authority" to mean "who bosses whom" and represent this by the familiar relation "greater than". If individual B commands a platoon and individual A commands the company which includes that platoon, then $A > B$. By such a means a simple order along a chain of command can be formed. There is some difficulty, however, about the definition of an equivalence class corresponding to a specific ordinal position, a problem at the level of the nominal scale. Consider two individuals, C and D , each commanding a separate group of twenty men. If there is a one to one correspondence between the members of one group and the members of the other such that the men are potentially interchangeable then C and D can be regarded as "bossing" the same men and hence members of an equivalence class. If the men are not interchangeable within pairs, however, the authority structure becomes a partial order again. Let us imagine, however, for the sake of simplicity, that corresponding elements are interchangeable. We have then a simple order of "authority".

The Ordered Metric Scale

In the three scales discussed heretofore—nominal, partially ordered, and ordinal—the elements of the system were classes of objects and the relationships were relationships of *equality* and *greater than*. It should be observed that nothing has been said about a concept of *distance* between classes. The formal introduction of distance by means of a distance function is beyond the

scope of this chapter, and the concept of distance between classes will be dealt with here on a purely intuitive level. In the scales discussed so far, all relationships such as "greater than" were between objects, which were stimuli or individuals. Consequently there was a complete absence of such a relationship between any pair of *distances* between objects. Thus, although A may have been observed to be greater than B , and B greater than C , nothing has been introduced concerning whether A was greater than B by a larger amount than B was greater than C . From this point of view the scales introduced up to this point may be regarded as nominal scales with respect to distances between classes of objects.

This introduces the concept of composite scales in which there are relations between the objects and in which, in addition, there are relations between the distances between objects. The nominal scale first discussed may be seen to be a nominal nominal scale, for it refers to objects first and distances between them secondly. Similarly, the partially ordered scale is a partially ordered nominal scale, and the ordinal scale is an ordered nominal scale.

It immediately becomes evident for any one of these scales that if, in addition, a relationship is observed to hold on the distances between classes then a higher level of measurement is achieved. When a relation of "greater than" holds for *some* pairs of distances between adjacent objects on an ordinal scale, the scale is an ordinal partially ordered scale. And if this relation holds for all pairs of such distances the scale is an ordered ordered scale.

For the sake of simplicity of discussion we shall not distinguish between those two levels but shall refer to them together as an ordered metric scale. By an ordered metric is meant a scale of which it can be said of any triplet of classes that $a > b > c$ and also that for at least some intervals between classes, e.g., the intervals \overline{ab} , \overline{bc} , \overline{ij} , \overline{kl} , either $\overline{ij} > \overline{kl}$ or $\overline{kl} > \overline{ij}$, where in general, \overline{jk} signifies the distance from j to k .

To illustrate the idea of an ordered metric we shall take the previous example of a simply ordered scale of "authority" and build it into an ordered metric scale. For convenience of discussion, it will be best to give names to the equivalence classes on the ordinal scale. Let these be in order of increasing authority, private corporal, buck sergeant and master sergeant. To establish an

ordered metric, there must be information in the data which leads to such conclusions as that the difference in authority between a corporal and a buck sergeant is either greater than, less than, or equal to the difference in authority between a private and a corporal. If the raw data consist of "who bosses whom," it might be possible to construct an ordered metric rationally. Suppose, for example, that a buck sergeant had command over two corporals each of whom commanded twenty privates. The buck sergeant would exceed the authority of a corporal by two corporals and twenty privates. Now consider the step from a buck sergeant to a master sergeant. Suppose that the master sergeant had command over three buck sergeants, each of whom commanded two corporals who commanded twenty privates each. The master sergeant would then exceed the authority of a buck sergeant by eighty privates, by four corporals, and by three buck sergeants. With respect to every element of authority, as defined here, the master sergeant then exceeds the buck sergeant by more than the buck sergeant exceeds the corporal, and the conclusion could be drawn that the increment in authority from corporal to buck sergeant is less than the increment from buck sergeant to master sergeant.

In order to construct this ordered metric scale of authority, certain assumptions had to be made: all squads of twenty privates were interchangeable, all corporals were interchangeable, and all buck sergeants were interchangeable. Thus, within each class, there were one or more units of measurement but they were not converted into a common unit. If the consequences of such assumptions lead to an ordered metric which is "unreasonable," as this one may well be, one would tend to regard the assumptions as bad and to seek other assumptions. The important point here is that these assumptions, whether implicit or explicit, recognized or not recognized, constitute part of the operational definition of "authority" in that they lead directly to a "measure" of "authority" and hence determine part of the meaning that the concept "authority" has when so measured.

Once an ordered metric scale of "authority" has been constructed in one way, it would be interesting to observe whether people perceive the metric relations as constructed here or whether, perhaps for psychological reasons, people perceive different metric relations. By use of the Method of Similarities, an adaptation of

scope of this chapter, and the concept of distance between classes will be dealt with here on a purely intuitive level. In the scales discussed so far, all relationships such as "greater than" were between objects, which were stimuli or individuals. Consequently there was a complete absence of such a relationship between any pair of *distances* between objects. Thus, although A may have been observed to be greater than B , and B greater than C , nothing has been introduced concerning whether A was greater than B by a larger amount than B was greater than C . From this point of view the scales introduced up to this point may be regarded as nominal scales with respect to distances between classes of objects.

This introduces the concept of composite scales in which there are relations between the objects and in which, in addition, there are relations between the distances between objects. The nominal scale first discussed may be seen to be a nominal nominal scale, for it refers to objects first and distances between them secondly. Similarly, the partially ordered scale is a partially ordered nominal scale, and the ordinal scale is an ordered nominal scale.

It immediately becomes evident for any one of these scales that if, in addition, a relationship is observed to hold on the distances between classes, then a higher level of measurement is achieved. When a relation of "greater than" holds for *some* pairs of distances between adjacent objects on an ordinal scale, the scale is an ordinal partially ordered scale. And if this relation holds for all pairs of such distances, the scale is an ordered ordered scale.

For the sake of simplicity of discussion we shall not distinguish between those two levels but shall refer to them together as an ordered metric scale. By an ordered metric is meant a scale of which it can be said of any triplet of classes that $a > b > c$ and also that for at least some intervals between classes, e.g., the intervals \overline{ab} , \overline{bc} , \overline{ij} , \overline{kl} , either $\overline{ij} > \overline{kl}$ or $\overline{kl} > \overline{ij}$, where in general, \overline{jk} signifies the distance from j to k .

To illustrate the idea of an ordered metric we shall take the previous example of a simply ordered scale of "authority" and build it into an ordered metric scale. For convenience of discussion it will be best to give names to the equivalence classes on the ordinal scale. Let these be, in order of increasing authority, private, corporal, buck sergeant and master sergeant. To establish an

(\overline{CD}), each larger than the step between corporal and buck sergeant (\overline{BC}), were, *in these data*, incomparable. The simplified technique used did not yield data which contained this information, hence no conclusion can be drawn.

In the process of constructing an ordered metric scale for 'authority,' we have actually constructed *two* of them, one by definition and the other as "perceived." The question naturally arises as to which one is "better" or "right." This gives rise to a basic question: What is "authority," anyway, and, for that matter, just what is an "attribute"? It will not be profitable to pursue this subject here, instead it will be sufficient to point out that this is where the doctrine of operationism plays its role. The concept of 'authority' has precisely such meaning as resides in the operations involved in observing it. To endow either of these scales with all the meanings and implications associated with the concept of "authority" is operationism in reverse and therefore specious.

To summarize, the types of scales which have been discussed are the nominal scale, the partially ordered scale, the ordinal scale, and the ordered metric. In this order they represent successively more powerful levels of measurement, in the sense that data which satisfy successive levels contain more and more information.

The Interval Scale

The level of measurement represented by the interval scale³ involves a step above the ordered metric corresponding to a considerable increase in "power." It will be recalled that the ordered metric was characterized by a simple ordering of the stimuli on a scale and by at least a partial ordering on the magnitudes of the distances between adjacent stimuli on the scale. The interval scale is characterized by the fact that the data contain information on just *how large* the intervals between all stimuli on the scale are. This requires a distance function which assigns a real number to all pairs of elements in an ordered set. Operationally this condition is satisfied by the existence of a common and constant unit of measurement. In such a case numbers may be associated with the positions of the stimuli on the scale such that the operations of

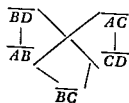
³ S. S. Stevens was the first to distinguish between and name the interval and ratio scales (29). See also Stevens (30).

the Unfolding Technique (6) (discussed in the last two sections), it is possible to take a single individual subject and determine the 'structure' of the attribute 'authority' as he perceives it for these stimuli. It can readily be determined whether his perception of these ranks satisfies a simply ordered system and what some of the metric relations are

Let A = private
 B = corporal
 C = buck sergeant
 D = master sergeant

To illustrate the concept of an ordered metric scale, these four stimuli were administered by the Method of Similarities to two individuals and each individual's structure of the concept of "authority" was obtained. The stimuli were presented to the subject three at a time and he was asked to judge of each triad which two were most nearly the same in authority and which two were least alike in authority. These data could be unfolded for each individual and certain characteristics of the stimulus space determined. In this instance, the scale of "authority" for those four stimuli was unidimensional for each of the two individuals who were tested, and the rank order was A, B, C, D . This, of course, is hardly surprising.

The information in these data on metric relationships between stimuli can be represented by the following partial order, which was the same for both individuals



where in general, \overline{JK} signifies the distance from stimulus J to stimulus K , and a line connecting two elements signifies that the one above is the larger. We see that the increment in 'authority' in going from corporal to buck sergeant (\overline{BC}) was psychologically the smallest of all. The increments represented by going from a private to a corporal (\overline{AB}) and from a buck sergeant to a master sergeant

(\overline{CD}), each larger than the step between corporal and buck sergeant (\overline{BC}), were, *in these data*, incomparable. The simplified technique used did not yield data which contained this information; hence, no conclusion can be drawn.

In the process of constructing an ordered metric scale for "authority," we have actually constructed *two* of them, one by definition and the other as "perceived." The question naturally arises as to which one is "better" or "right." This gives rise to a basic question: What is "authority," anyway, and, for that matter, just what is an "attribute"? It will not be profitable to pursue this subject here; instead it will be sufficient to point out that this is where the doctrine of operationism plays its role. The concept of "authority" has precisely such meaning as resides in the operations involved in observing it. To endow either of these scales with all the meanings and implications associated with the concept of "authority" is operationism in reverse and therefore specious.

To summarize, the types of scales which have been discussed are the nominal scale, the partially ordered scale, the ordinal scale, and the ordered metric. In this order they represent successively more powerful levels of measurement, in the sense that data which satisfy successive levels contain more and more information.

The Interval Scale

arithmetic may be meaningfully performed on the *differences* between these numbers

It will be easiest to illustrate the concept of an interval scale by returning to the ordered metric previously constructed for 'authority' and to try to convert it into an interval scale. To do this, a decision must be made as to what shall constitute a common and constant unit of authority for all classes or ranks. Suppose one decided that command over one private constituted a unit of authority, and that a corporal would thus represent twenty units of authority. A buck sergeant who commanded two corporals each commanding twenty men would, on this basis, be regarded as having eighty units of authority. A master sergeant, commanding three buck sergeants and all their men, would then represent 480 units of authority (120 privates at one unit each, six corporals at twenty units each, and three buck sergeants at eighty units each).

The scale scores of those grades or ranks could then be represented as follows

<i>Scale score⁴</i>	<i>Rank</i>
1	private
20	corporal
80	buck sergeant
480	master sergeant

As this is an interval scale differences between scale values may be operated on arithmetically. Thus, for example, it can be said that a buck sergeant has sixty more units of authority than a corporal, a master sergeant has 400 more units of authority than a buck sergeant, and consequently the latter increment in authority is $6\frac{2}{3}$ times the increment from a corporal to a buck sergeant.

A significant aspect of an interval scale is that the numbers associated with the points on the scale are 'right' only within a linear transformation. For example, any given number can be *added*

⁴ This again is an example of how a concept is endowed with meaning by measuring it. There are other definitions which could be made which would result in nonlinear transformations of these measures. For example one might argue that a buck sergeant commands only the corporals and not the men under the corporals and similarly that the master sergeant commands only the buck sergeants. The scale scores would then be 1 20 40 120. Then if authority so measured bore a linear relation to some other variable authority as measured in the text above would have a nonlinear relation to that variable.

to all the scale scores and the *relations between intervals* will be preserved. This is called "translation" and corresponds to the fact that the origin, the location of the numerical value "zero," is arbitrary. Similarly all the scale scores may be *multiplied* by any given number and the relations between intervals will be preserved. This is called "scalar multiplication" and corresponds to the fact that the unit of measurement is arbitrary. A linear transformation consisting of both a shift in origin and a change in the unit of measurement will also preserve the relations between intervals.

This is an important consideration if one wants to compare the authority of a sergeant with, for example, the authority of a foreman of a repair crew. In the latter case, the unit of authority might be chosen to be "one repairman." To be meaningful, the comparison then requires that the origin, zero, be the same in both scales and that the unit of measurement be identical. This is the reason why one cannot compute the significance of a difference between the mean height and the mean weight of a group of people. As numbers are being used in each case, there is a tendency to apply the operations of arithmetic because this is possible and easy to do in the abstract system of numbers; but the consequences of the operations on the numbers do not necessarily have meaning in the real system of objects which the numbers represent. For such a test of significance to be valid, *both scales* must be measured on the *same* interval scale. Under such conditions the interval scale permits the valid use of most of the tools of mathematics and statistics. It is interesting to note that the product moment correlation does not require that both variables have the same origin and unit of measurement because this index of relationship is invariant under independent linear transformations of the variables.

The Ratio Scale

To complete this discussion of the various levels of measurement, the ratio scale will be presented briefly. The ratio scale is an interval scale with the additional characteristic that the origin is an absolute zero and not an arbitrary zero. In such circumstances, the operations of arithmetic are permissible not merely on the differences, as was the case in the interval scale, but on the scale values themselves. The numbers associated with scale values on a ratio

scale are then right within scalar multiplication, a consequence of the fact that only the unit of measurement is arbitrary. Under these conditions, it is possible to compute a meaningful ratio of two scale values. Suppose, for example, in the case of the interval scale of authority previously constructed, that zero could be properly associated with the scale value of a prisoner of war⁵ as an absolute zero of authority. Then the scale values can no longer be translated but can only be multiplied by a scalar. The buck sergeant can then be said to have four times the authority of a corporal, and a master sergeant to have 24 times as much authority as a corporal. These relations would of course, be preserved under scalar multiplication.

Measurement vs. Scaling

Because of the considerable increment in power between an ordered metric and an interval scale, there is a tendency to distinguish between two broad classes of scales. The theory of the ordered metric and less powerful scales may be referred to as scaling theory, and the theory of the interval and ratio scales as measurement theory (7). The former may also be thought of as qualitative measurement (if this is not a contradiction in terms) and the latter as quantitative measurement. Because of the composite character of the scales with respect to their logical structure, there is also a natural tendency to refer to the entire domain as measurement theory.

From the previous discussion of levels of measurement, it should now be apparent that there are two broad aspects to measurement. On the one hand there is an abstract or formal system of elements with certain properties and operations. Each such system is a mathematical system, a calculus in the sense in which Carnap uses it (4). The successive levels of measurement correspond to successive systems in which there are, axiomatically, more properties and operations. In these cases the result is a successively stronger system in the sense of the proliferation of the mathematical structure that can be built on the given axiomatic basis.

The second broad aspect of measurement consists of the do

⁵ This of course is not experimentally demonstrable by any existing technique and could be accomplished only by assumption.

mains of "objects" in the real world and their observable properties and relations. Measurement may then be regarded as the process of mapping a real object system into one of these abstract systems, the purpose of the mapping being, in part, to substitute operations in the abstract system for operations on the object system. However, in order that the operations in the abstract system have meaning in terms of the real object system, it is necessary that the axioms on which the abstract system has been built be satisfied by the object system.

To say that measurement may be regarded as mapping an object system into an abstract system is both more and less than what is sometimes meant by measurement. From a limited to a general sense measurement may mean

- (1) mapping an object system into an interval or ratio scale, permitting the assignment of numbers to objects, and permitting at least some of the operations of arithmetic to be performed on these numbers
- (2) mapping an object system into at least a simple order, including the ordered metric, interval scale, and ratio scale
- (3) a generalization to the extent of mapping an object system into a partial order or even a nominal scale
- (4) a generalization including the decomposition of partial orders into sets of simple orders—in effect, measurement in a multidimensional space

Under this last generalization, measurement theory includes the entire subject of analysis of data. This is as it should be, for data analysis is an integral part of the logic of measurement theory.

The Dilemma of the Social Scientist

It is now evident that the process of measurement consists in part of selecting a level of measurement, an abstract system, into which the data are to be mapped. A variety of such systems is available, varying in level from weak to strong. The axiomatic basis of the level of measurement selected constitutes a theory about the behavior in question in that the axioms specify the relationships that are to hold in the data and the properties of the relationship. These axioms then become part of the data in the sense that they take

precedence over the data. If there should exist any data incompatible with the axioms of the level of measurement selected, these data constitute error, by definition.

Let us illustrate this by a hypothetical example of rating leadership ability. Suppose we have a group of individuals whose leadership abilities it is desired to determine, and suppose a group of judges who know each of the individuals are asked to rank order them in leadership ability. We shall require that leadership ability be measured on an ordinal scale. This requires that if A has more leadership ability than B , and B has more than C , A must have more than C , for all triplets, A, B, C . Inasmuch as we required each judge to rank order the individuals, our method of collecting data has imposed a simple order on leadership ability as far as each judge is concerned. But we find that the judges do not all agree on the rank order of the individuals in leadership ability. Consequently we must make some further assumptions, usually pertaining to randomness of errors, a constant origin, and a constant unit of measurement. Such assumptions as these will make possible the combination of a number of different rank orders into the single rank order which is required. One may then return to the data and specify in detail the errors which each judge made. It would perhaps be interesting to face the individual judges with their alleged errors and study their reactions.

The problem implicit in the situation can be called the dilemma of the social scientist—which in its simplest form is the problem of what shall be called error. Almost anyone is willing to say that any given set of data contains some error, but just what is to be classified as error depends a good deal on the level of measurement assumed to hold in the data.

The social scientist is faced by his dilemma when he chooses between *mapping* his data into a simple order and *asking* his data whether they satisfy a simple order. By selecting a strong enough system, the social scientist can always succeed in constructing a unidimensional scale of measurement, commonly an interval scale, thus requiring a portion of the data to be classified as error. By not *requiring* a strong system the social scientist permits the data to determine whether a simple unidimensional solution is adequate. Unidimensionality, obtained by a method of analysis which guaran-

tees it, obviously cannot thereby be shown to be a characteristic of the behavior in question. This is merely a special case of a more general principle that no property of data can be said to hold unless the methods of collecting and of analyzing the data permit alternative properties to exhibit themselves. The problem of the social scientist, in blunt terms, is whether he knows what he wants or whether he wants to know.

There are several reasons why social scientists so frequently choose a strong level of measurement rather than a weak one to represent the data. It is always more profitable to use a strong scale in preference to a weak one when both satisfy the data, because a more powerful mathematics is available for use in the description and analysis of the data. Compare, for example, the rank-order correlation methods appropriate to an ordinal scale with the powerful systems of linear, multiple, and nonlinear correlation methods which require an interval scale. There is no wonder that measurement by an interval scale has been a major objective.

There is another, and in some ways a curious but valid, reason for the social scientist to choose a stronger level of measurement than is satisfied by the data—society often requires that at least a simple order be imposed on an attribute. Thus, in the case of the esthetic merit of paintings, an individual may be faced with the problem of choosing a painting to buy. In order to make a decision, he must, in spite of the existence of only a partial order, impose a simple order over at least a segment of the space. This conflict between what appears to be the inherent nature of social-psychological attributes as revealed by the data and the common insistence by society that at least a simple order be imposed lies at the root of the problem which the social scientist faces. This is the primary explanation of why the social scientist must so frequently be "unscientific" and, in effect, be forced to treat his measurement theory or scale as "right" in spite of his data.

This situation, the common social necessity of mapping a partial order into a simple order or stronger system, has no unique solution except by fiat. There may be a number of different "best" solutions for different purposes, but there appears to be no single universally "right" solution. This enforced mapping of a partial order into a stronger system may be one of the sources of social conflict.

The major implication to be drawn from the discussion in this section is that it behooves the social scientist to become fully aware of and acquainted with the subject of measurement theory. There are no pat answers or conventional rules which can be applied routinely and in the absence of understanding. The "conclusions" that are drawn from an analysis of data are highly dependent on the level of measurement assumed and imposed on the data, 'conclusions' may be easily confused with postulates or assumptions that have been built into the data. A decision must be made in terms of the objectives of the specific study as well as on the basis of the logic of measurement theory.

This section has discussed the general nature of measurement theory, the various levels of measurement, and what is involved in the process of measuring. These considerations, combined with the theory of data presented in the next section, will provide a framework for understanding and relating the various methods of collecting and analyzing data discussed in the last two sections.

A THEORY OF DATA

There is a great variety of techniques and procedures used by social scientists for collecting and analyzing data. All too frequently, studies which should have mutual implications are not comparable because their methodologies are different. The purposes of the theory of data presented in this section are to provide a simple framework for organizing and classifying the methods of collecting and analyzing data and to provide a mathematical basis for relating methodologies within classes and between classes.

Genotypic and Phenotypic Levels of Description

In the theory of data presented here, two levels of description will be recognized—a genotypic level and a phenotypic level. The phenotypic level refers to the observed or manifest behavior, the genotypic level to an inferred, hypothetical, latent level of behavior underlying or generating the phenotypic level.

These two levels of description can be illustrated by the per

formance of an individual on a mental test. On an arithmetic test, for example, his behavior on each item (usually categorized as right or wrong) constitutes the manifest behavior, the phenotypic level. On this level, performance is commonly represented by a pattern of responses—passes and failures—or by a numerical score based on the number of items right. From the manifest behavior of this individual and of a number of other individuals inferences of a genotypic nature are usually made—for example, that individual *A* has more of the ability than individual *B*. The manifest behavior is implicitly regarded as a function of the individual's genotypic ability and certain characteristics of the stimulus situation. On another test, also measuring arithmetic ability, the individual may get a quite different score, but a set of such scores from a number of individuals are conventionally expected to have at least a monotonic, if not linear, relationship to the scores of these same individuals on the first test.

In a different context, suppose that the more aggressive of two individuals always dominated. A given individual with a particular amount of aggressiveness (genotypic level) might behave submissively (phenotypic level) in the presence of one individual and be dominant in the presence of another. Some of the assumptions involved in such a line of reasoning have been formalized and made a part of a basis for a general theory of psychological scaling (8). The theory of data presented here, and all of the relationships between methodologies contained in the next two sections, are a consequence of the general theory. In this chapter the more technical treatment will be avoided and the presentation will be made on a verbal level.

The Information Contained in Observations

There are two fundamental aspects to this theory of data—a definition of the information contained in an observation on the phenotypic level, and a definition of the relationship between the genotypic level and the phenotypic level. These two aspects provide the basis for making genotypic inferences from the observations.

Let us consider what the information is that is contained in an individual's performance on an arithmetic item. Genotypically,

the stimulus⁶ an arithmetic problem in this case, is regarded as requiring a certain minimum amount of some ability for it to be successfully solved or passed (the ability is spoken of in the singular but it may be constituted of more primitive abilities) The individual is regarded as possessing a certain amount of this ability His manifest behavior consists in either passing or failing the problem The assumption is made that this implies that the individual has more or less, respectively of the ability required⁷ The information in each element of behavior, a response to an item in the test, is whether the individual has more or less of some ability than the amount of the ability required by that item

This is the kind and the amount of information which will be assumed to be contained in the behavior of individuals in responding to this type of item Exactly what this type of item is will be discussed later when these ideas are generalized

There are other ways of collecting or observing behavior and other types of items These procedures may differ both in the kind and in the degree of information they contain Consider, for example, the manifest behavior of an individual in stating whether or not he will endorse certain opinions A statement of opinion may be regarded as representing a particular degree of attitude and possibly different degrees on different attitudes On each of such attitudes, the individual has a specific degree of attitude himself at any moment of time The degree of attitude held by an individual will be called his *ideal* for that particular attitude Whether or not an individual endorses a given item is then determined by whether he regards the item as being 'sufficiently close' to his ideal If the individual is asked to choose the three items he most prefers to endorse, it will be assumed that he chooses the three items nearest his ideal

It is apparent that the information contained in such data is distinct from the kind of information contained in the example of the arithmetic test In the arithmetic test, the individual 'passed' all items requiring less ability than he possessed, in the

⁶ Whether behavior arises in response to other individuals group situations internal physiological states or whatever these may each be regarded as a stimulus situation For simplicity it will be called a stimulus

⁷ The items being discussed do not permit chance success by guessing in other words they are not objective type items

case of the attitude items, he 'endorsed' those nearest to him. To generalize these ideas somewhat, the individual's ability may also be regarded as an 'ideal'—he passed all the arithmetic items on one side of his ideal and failed all those on the other side. In the case of the attitude items, he endorsed items near his ideal regardless of direction and rejected items further away.

In these two instances, the manifest behavior is abstractly the same. The stimuli are classified for each individual in two piles: those he passed or endorsed and those he failed or rejected. The assumptions by means of which genotypic inferences are made, however, are distinct in these two examples.

With some methods of collecting data, the information in the manifest behavior differs in degree from the above. For example, the individual may be asked to rank the statements of opinion in the order in which he prefers to endorse them. On the basis of the assumptions already made, it follows that the stimuli are ranked in order of their increasing distance from his ideal on the respective attributes.

It is obvious that the information in these data differs in kind from the information in data based on which items the individual is actually willing to endorse. In the rank order data it is not known whether the individual would be willing to endorse any of the items or all of them. It is also apparent that the information in the rank order data differs only in degree from that contained in data based on which *three* items the individual would prefer to endorse. The *three* items the individual would choose are, on the basis of the postulates, the first three items in the rank order.

Task A and Task B

The kinds of behavior which have been discussed up to this point have one thing in common from the point of view of the theory of data. They all involve the evaluation of stimuli with respect to an ideal. In each instance both the stimuli and the individuals have hypothetical genotypic measures and the manifest behavior permits inferences to be made about the relationship of the genotypic magnitudes of the stimuli to those of the individuals. In the later sections of this chapter, particularly in the last section, we shall see what conditions are necessary to make inferences about the geno-

typic relationships between individuals and the genotypic relationships between stimuli

For didactic purposes data involving the evaluation of stimuli with respect to an ideal will be referred to as data collected by task *A*. All the behavior discussed up to this point, then, is task *A* behavior. It should perhaps be pointed out that it makes no difference whether an attribute is explicit or implicit in the instructions to the subject. Thus it is still task *A* whether an individual is asked which candidate he prefers with respect to their attitudes toward foreign affairs or whether he is merely asked which candidate he prefers. This does have a bearing on the interpretations given to the genotypic inferences but in the formal analysis of data it is irrelevant. Similarly, it makes no difference in the abstract system whether the individual's ideal is his own, someone else's or one given to him by the experimenter. The abstract characteristic of task *A* is that a stimulus is evaluated with respect to direction or distance⁸ from a point in space called an ideal.

There is another kind of behavior which can be observed and with respect to which data can be collected. This type of behavior is evaluation of stimuli with respect to an attribute and will be called task *B*. Illustrations are contained in the evaluations of statements of opinion as to which candidate expresses a more liberal attitude or which candidate is more pro union or in rating the aggressiveness of an individual. Whether or not the judge has an ideal of his own on liberalism etc., is regarded as irrelevant.

Data collected under task *B* may also differ both in kind and degree. When an individual ranks a number of individuals as to their administrative capacities, there is no information in these data as to whether or not any of the individuals are *good* administrators as would be implied in the case of rating them. This is an example of a difference in *kind* of information. An example of a difference in *degree* of information is picking the three best administrators in a group of individuals; this contains less information than the rank ordering.

⁸ In a multidimensional space the definition of distance may be made in several ways. This problem is too complex and has not been sufficiently worked out for discussion in this chapter. In the unidimensional case these alternative definitions of distance do not arise.

Relative and Irrelative Behavior

In this discussion of the theory of data, an explicit dichotomy has been made between task *A* and task *B* with respect to the nature of the information contained in the data. Certain other dichotomies have been made implicitly. One of these may be referred to as relative and irrelative behavior. In relative behavior, the data are based on the relationships between two or more stimuli—for example a judgment as to which of two candidates an individual would prefer (task *A*), or which of the candidates is more pro union (task *B*). If the individual rank ordered his preference for all the candidates it would still be relative behavior, task *A*, and the information in

	IRRELATIVE BEHAVIOR	RELATIVE BEHAVIOR
TASK A	<div> <div>Monotone Stimuli Nonmonotone Stimuli</div> <div>IIa IIb</div> </div>	I
TASK B	III	IV

FIG. 1 A classification of methods of collecting and analyzing data

the data would differ only in degree from the information in a judgment on one pair. Obviously this can be extended to task *B* similarly.

In irrelative behavior, the individual's judgments involve a single stimulus at a time. This is illustrated by the arithmetic test referred to by rating individuals on a rating scale, or by expressing one's likes and dislikes for each of a number of items in an interest inventory.

The two dichotomies developed so far, tasks *A* and *B*, and relative and irrelative behavior, may be put together in the form of a fourfold table (Fig. 1) in which for simplicity of discussion, the

quadrants have been numbered. It will be observed that Quadrant II has been further dichotomized on the basis of the kind of stimuli: monotone or nonmonotone. This distinction is illustrated by the different assumptions involved in making genotypic inferences in the passing or failing of an arithmetic item and in the case of whether or not a statement of opinion is endorsed. In the arithmetic items the individual's manifest behavior is *pass* for all items on one side of him and *fail* for all items on the other side. This type of item is called *monotone*. Nonmonotone items are exemplified by statements of opinion in which the individual endorses those items nearest him in a segment of the space surrounding him and rejects all items beyond.

In the case of a single underlying latent attribute the items an individual rejects may consist of two subsets: one of which contains those that are too extreme in one direction and the other those that are too extreme in the other direction. Consider, as an example of a nonmonotone item, the following statement of opinion: "We should make the loan to Britain if we are sure they will pay it back." On a hypothetical continuum from *pro* to *anti* British individuals in the neighborhood of the middle of this continuum would presumably endorse this statement. But the individuals who refuse to endorse this statement may be at opposite poles of the continuum. The very *pro* British may say *no* because they want to make a loan to Britain without any conditions and the very *anti* British say *no* because they do not want to make a loan to Britain under any conditions. Hence two very different kinds of people genotypically behave identically (phenotypically) in this situation. Similarly those individuals in the middle of the continuum who endorsed this statement might be expected not to endorse extreme *pro* or *anti* British statements. Hence certain individuals will behave phenotypically the same to two genotypically distinctly different kinds of stimuli.

In the abstract generalization of these two types of items, *monotone*⁹ items are those for which a one to one mapping of the categories of manifest behavior into genotypic categories is possible and nonmonotone items are those for which a one to many mapping is necessary.

⁹ A type of item called *cumulative* is a special case of a monotone item.

Distinction Between Methods of Collecting and Analyzing Data

We now have a framework within which methods of collecting and analyzing data may be discussed. Within a given quadrant (Fig 1), the *kind* of information contained in the data is the same and data collected by various methods differ primarily in the degree of information they contain. The placement of a set of data in this fourfold table will be seen to be, within certain limits, a decision that is made by the experimenter when he chooses a particular technique for analyzing the data.

The method of collecting data *determines* what information they contain, but the method of analysis *defines* this information, and this is what situates the data in the table. Methods of analyzing data have been devised historically, *ad hoc*, for each of the quadrants. Some of these methods seek only to consolidate or 'average' the phenotypic information, others seek to make genotypic inferences from the phenotypic information.

The method of analysis selected may permit the discovery of the properties of the information or may also *define* the properties. In the latter case, the experimenter is concerned only with the interrelations. This is precisely the 'locus of the dilemma of the social scientist' referred to in the previous section.

The distinction between methods of collecting data and methods of analyzing data is imperative for an understanding of the relationship between the inferences drawn from different studies. The relationships among the quadrants are basic to an understanding of the distinction between collecting and analyzing data and also to the relationships among different methods of collecting data and different methods of analyzing data.

Irrrelative behavior (represented by Quadrants II and III) is, in the abstract, the response of an individual to stimuli *per se* as contrasted with relative behavior (Quadrants I and IV), in which the response of an individual is a *choice* between stimuli. Clearly irrelative behavior is represented by the Method of Single Stimuli in its broadest sense, and all the methods of collecting data in Quadrants I and IV, relative behavior, can be referred to collectively as the Method of Choice.

In Quadrant III a distinction is possible between monotone

and nonmonotone stimuli, but since methods of collecting data in this quadrant invariably use monotone stimuli, nonmonotone stimuli have been neglected in this presentation. It can be shown that Quadrant III is formally indistinguishable from Quadrant IIa—the distinction exists only in the frame of mind of the experimenter. Furthermore, it can be shown that the Method of Single Stimuli as a whole, both Quadrants II and III, is a special case of Quadrant I.

The distinction between Quadrants I and IV is a real one although there are data collected by certain methods which may be placed in either quadrant depending upon the objectives and frame of mind of the experimenter. These distinctions and interrelations will be brought out in more detail in the next two sections.

METHODS OF COLLECTING DATA

In this section a number of methods of collecting data will be discussed and some of their interrelations pointed out. A potential source of confusion resides in the fact that the names for some methods imply both a method of collecting and a method of analyzing data—e.g., the Method of Successive Intervals (28). Throughout this section the mention of any method will have reference only to its use as a method of *collecting* data.

A general system for structuring or organizing methods of collecting data in Quadrant I will be developed on the basis of the *amount* of information each contains. The relation of these methods when used to observe behavior in Quadrant IV will be pointed out. The sense in which the Method of Single Stimuli of Quadrants II and III is a special case of Quadrant I will then be discussed.

Quadrant I

To illustrate the information contained in methods of collecting data in Quadrant I, examples will be given in which unidimensionality has been imposed. Such examples will be used partly because the unidimensional case has been more completely worked out and also because of the preoccupation of most social scientists with unidimensional representations of data.

Let us suppose, throughout the following discussion that we have five stimuli, *A*, *B*, *C*, *D*, and *E*, on some latent attribute and that the ideals of individuals are distributed over this same latent attribute with fixed metric relations. For concreteness imagine that the five stimuli are statements of opinion representing different degrees of attitude toward *x*, and the ideals of the individuals are the hypothetical statements of opinion which each would prefer to endorse above all others. This situation represents what in general are unrealistic constraints, but the simplicity of the conditions is very desirable for didactic purposes. These constraints are completely relaxed in the more general treatment.

The situation may be illustrated by Figure 2, in which a frequency distribution of the ideals of the individuals over a latent

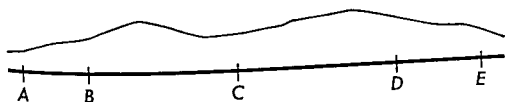


FIG 2 An example of a distribution of people and stimuli on a Joint scale

continuum is indicated. Such a scale may be called a *J* scale, for joint scale or joint distribution, having both stimuli and individuals on it.

Suppose that one collects data by asking each individual to indicate the two statements he most prefers to endorse. On the basis of the postulate that an individual will prefer to endorse a stimulus nearer his ideal than one farther away, the response patterns under these conditions would be the following:

AB, BC, CD, DE

It is evident that all individuals to the left of the midpoint between stimuli *A* and *C* would give the response *AB*. Those individuals to the right of the midpoint *AC* and to the left of the midpoint *BD* would have the response *BC*. The relationship of all responses to midpoints is illustrated in Figure 3, in which the boundaries of the regions associated with a response are indicated by vertical lines sectioning the scale. The response associated with a region or segment of the scale is also indicated.

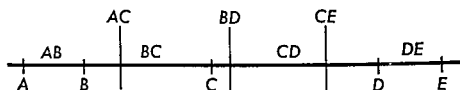


FIG 3 Relation of response patterns and scale values under
Pick 2

This method of collecting data called Pick 2 in general yields $n-1$ categories of individuals where n is the number of stimuli and the boundaries of regions are given by the midpoints of alternate stimuli

If instead of Pick 2 the individuals had been instructed to Pick 3 the response patterns would have been

ABC BCD CDE

and their relation to the scale would be as illustrated in Figure 4

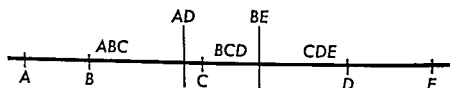


FIG 4 Relation of response patterns and scale values under
Pick 3

The number of categories of individuals would have been three or in general $n-2$. This procedure Pick k may be continued for higher values of k up to $n-1$. For Pick $n-1$ only two patterns of response would be obtained. One pattern would be that of individuals to the left of the midpoint between the first and last stimulus the other pattern would be that of individuals to the right of this midpoint. This might be more immediately obvious if one recognizes that from a formal point of view Pick $n-1$ is the same as Reject 1. In the unidimensional case there are only two stimuli which may be rejected the two end ones and the choice between these two is dependent on their midpoint in relation to the judges' ideal. In the present example the results would be as indicated in Figure 5

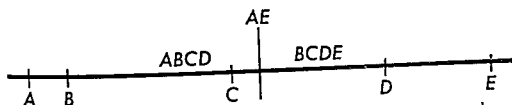


FIG. 5. Relation of response patterns and scale values under "Pick $n-1$ " or "Reject 1."

For each of these methods of collecting data in Quadrant I, from "Pick 2" to "Pick $n-1$," there corresponds a method of analysis which will reveal certain characteristics of this genotypic structure underlying the phenotypic behavior—the stated preferences. This method of analysis, called Parallelogram Analysis, is discussed in the next section.

If this method of collecting data, "Pick k ," were extended to $k = 1$, where each individual indicates the one item he will endorse, there is no unique solution to the genotypic structure underlying the preferences (see p. 513). "Pick 1" is the special case of the Method of Choice which corresponds to the Method of Single Stimuli. This is not so severe a criticism as it may at first appear, because the Method of Single Stimuli is ordinarily used only where there is an *a priori* ordering of the alternatives to an item and the *data* are not asked to provide it. Methods of analyzing such data then become concerned only with the interrelationships between a number of such scales as the above, one for each item. Consequently, "Pick 1" will be studied as a special case in its own right, under Quadrant II. Omitting "Pick 1," then, for the time being, let us return to "Pick 2" and continue with relative behavior.

If our subjects are instructed to "Order 2" (*i.e.*, indicate their first and second choices) instead of "Pick 2," the response pattern AB obtained in "Pick 2" becomes two response patterns AB and BA and these are given by individuals to the left and right, respectively, of the midpoint AB . Similarly, each of the "Pick 2" patterns becomes two patterns under "Order 2" instructions, and the midpoints of adjacent stimuli on the genotypic scale have been added to the midpoints of alternate stimuli to form the boundaries of the regions associated with each phenotypic response pattern. This is illustrated in Figure 6. The number of categories of individuals has become eight instead of four, or, in general, $2(n-1)$.

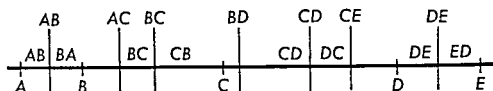


FIG 6 Relationship of response patterns and scale values under Order 2

It is now apparent that for n stimuli presented at a time, the methods of collecting data from 'Pick $n-1$ to Order 2' (omitting Pick 1) contain a monotonically increasing amount of information in the sense of the number of categories the individuals are classified into on the basis of distinguishable response patterns. The method of analysis of information (the Parallelogram Technique) leading to genotypic inferences is the same for all and will be discussed in the next section.

It is obvious that this series of methods of collecting data may be extended to 'Order 3,' Order 4, down to Order $n-1$. The latter corresponds to the Method of Rank Order. These methods of collecting data may be distinguished from the preceding methods for although both types are logically contiguous, methods where a number of items are ordered yield a new *class* of information—information about metric relations.

A detailed presentation of the theory and technique for extracting information about metric relations in rank order data is contained in the literature (6). Here the case of Order 3 will be illustrated. Keeping in mind Figure 2 and the psychological postulates which have been made, consider what the data would be like for 'Order 3.' All individuals to the left of the midpoint AB would yield ABC as their phenotypic behavior. Individuals to the right of this midpoint AB but to the left of the midpoint AC would act alike phenotypically and would yield BAC . Crossing the midpoint AC reverses the order of these two stimuli in the phenotypic behavior, and the next ordering would be BCA . If this process is continued for the situation given in Figure 2 the complete results are those given in Figure 7.

In comparison with 'Order 2,' where there were eight classes of individuals or, in general, $2(n-1)$, here there are ten or, in general, $3n-5$. But there is, in addition, a new kind of information in these

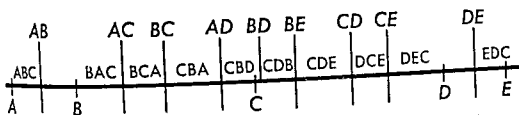


FIG. 7. Relation of response patterns and scale values under "Order 3"

data—information about metric relations. Consider the two midpoints BC and AD . In this instance the midpoint BC precedes AD because the interval between the two stimuli A and B , \overline{AB} , is less than the interval between the two stimuli C and D , \overline{CD} . The phenotypic behavior associated with the interval bounded by the two midpoints BC and AD is CBA , and this phenotypic behavior has the genotypic implication that $\overline{AB} < \overline{CD}$. If, for example, $\overline{AB} > \overline{CD}$ and hence the AD midpoint had preceded the BC midpoint, the phenotypic behavior in the region they bounded would have been BCD instead of CBA .

Similarly, the phenotypic behavior CDE implies that BE precedes CD , which has the genotypic implication that $\overline{BC} > \overline{DE}$. In general metric information from "Order 3" is obtained with respect to the relative magnitude of alternate single intervals between stimuli.

If "Order k " is employed as a method of collecting data, the information contained in the data includes that for any lower value of k and additional metric information as k increases. For $k \geq 4$, the information increases to include the relative magnitudes of sums of adjacent intervals compared with sums of adjacent intervals. The method containing the most metric information is "Order $n-1$," the Method of Rank Order.

For the sequence of methods of collecting data involving "Order k ," $3 \leq k \leq n-1$, there is a corresponding method of analysis for obtaining the genotypic inferences contained in the data. This method of analysis is called the Unfolding Technique. The rank order of the stimuli for an individual will be called a simply ordered I scale and may be regarded as the J scale folded at the individual's ideal. It is this which gives the name the Unfolding Technique to the analysis of sets of I scales (phenotypic behavior) to generate

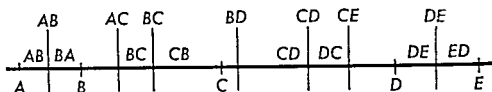


FIG 6 Relationship of response patterns and scale values under Order 2

It is now apparent that for n stimuli presented at a time the methods of collecting data from Pick $n-1$ to Order 2 (omitting Pick 1) contain a monotonically increasing amount of information in the sense of the number of categories the individuals are classified into on the basis of distinguishable response patterns. The method of analysis of information (the Parallelogram Technique) leading to genotypic inferences is the same for all and will be discussed in the next section.

It is obvious that this series of methods of collecting data may be extended to Order 3, Order 4, down to Order $n-1$. The latter corresponds to the Method of Rank Order. These methods of collecting data may be distinguished from the preceding methods for although both types are logically contiguous methods where a number of items are ordered yield a new *class* of information—information about metric relations.

A detailed presentation of the theory and technique for extracting information about metric relations in rank order data is contained in the literature (6). Here the case of Order 3 will be illustrated. Keeping in mind Figure 2 and the psychological postulates which have been made, consider what the data would be like for Order 3. All individuals to the left of the midpoint AB would yield ABC as their phenotypic behavior. Individuals to the right of this midpoint AB but to the left of the midpoint AC would act alike phenotypically and would yield BAC . Crossing the midpoint AC reverses the order of these two stimuli in the phenotypic behavior and the next ordering would be BCA . If this process is continued for the situation given in Figure 2, the complete results are those given in Figure 7.

In comparison with Order 2 where there were eight classes of individuals or in general $2(n-1)$ here there are ten or in general $3n-5$. But there is in addition a new kind of information in these

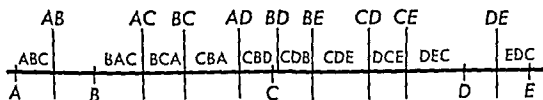


FIG 7 Relation of response patterns and scale values under Order 3

data—information about metric relations. Consider the two midpoints BC and AD . In this instance the midpoint BC precedes AD because the interval between the two stimuli A and B , \overline{AB} , is less than the interval between the two stimuli C and D , \overline{CD} . The phenotypic behavior associated with the interval bounded by the two midpoints BC and AD is CBA , and this phenotypic behavior has the genotypic implication that $\overline{AB} < \overline{CD}$. If, for example $\overline{AB} > \overline{CD}$ and hence the AD midpoint had preceded the BC midpoint the phenotypic behavior in the region they bounded would have been BCD instead of CBA .

Similarly, the phenotypic behavior CDE implies that BE precedes CD , which has the genotypic implication that $\overline{BC} > \overline{DE}$. In general metric information from 'Order 3' is obtained with respect to the relative magnitude of alternate single intervals between stimuli.

If 'Order k ' is employed as a method of collecting data the information contained in the data includes that for any lower value of k and additional metric information as k increases. For $k \geq 4$, the information increases to include the relative magnitudes of sums of adjacent intervals compared with sums of adjacent intervals. The method containing the most metric information is 'Order $n-1$ ', the Method of Rank Order.

For the sequence of methods of collecting data involving Order k , $3 \leq k \leq n-1$, there is a corresponding method of analysis for obtaining the genotypic inferences contained in the data. This method of analysis is called the Unfolding Technique. The rank order of the stimuli for an individual will be called a simply ordered I scale and may be regarded as the J scale folded at the individual's ideal. It is this which gives the name the Unfolding Technique to the analysis of sets of I scales (phenotypic behavior) to generate

a J scale (a genotypic inference) The Method of Parallelogram Analysis will be seen to be a special case of the Unfolding Technique

The next two methods of collecting data to be considered are the Method of Paired Comparisons and the Method of Triads The Method of Paired Comparisons for Quadrant I behavior constitutes the presentation of all possible pairs of stimuli with the instruction to the individual to indicate a preference within each pair The Method of Triads for Quadrant I behavior constitutes the presentation of all possible triplets of stimuli with the instruction to indicate which is most preferred and which is least preferred A further generalization is clearly possible but would be beyond the scope of this chapter

In the Method of Triads, each *pair* of stimuli is judged n^2 times, where n is the number of stimuli Each pair of stimuli is a constituent of a triad with each of the remaining stimuli in turn, and when the individual says of the stimuli A , B , and C that he prefers A most and C least, he is placing them in a rank order of preference $A > B, C$ This rank order of preference is equivalent to three transitive paired comparisons $A > B, B > C, A > C$, where $>$ means preferred It is important to note that this transitivity has been imposed by the method used in collecting the data The judgments of an individual on a triad may be decomposed into *three* paired comparisons which have *two* degrees of freedom One has no way of telling, of course, which two paired comparisons are the independent ones

Because each paired comparison judgment between a given pair of stimuli is made n^2 times, it is apparent that the Method of Triads permits the *consistency* of a paired comparison judgment to be tested in the context of a third stimulus The Method of Paired Comparisons, in which each paired comparison judgment is made only once, does not permit testing the *consistency* of a judgment but, assuming consistency, this method permits testing the *transitivity* of the paired comparison judgments It is now apparent that the Method of Rank Order imposes both consistency and transitivity on the implied paired comparison judgments

Putting together these various methods of collecting data, we have a simple order of methods from 'Pick $n-1$ ' to the Method of Triads, listed in order in Table I At the top of this "power" structure is the method among those discussed here that contains the

most information about the behavior being studied. The other methods of collecting data are listed in order of decreasing information. For convenience of discussion, it is desirable to give a name to the implied attribute underlying this power structure. The term "searchingness" is suggested. If we measure a unidimensional attribute, the methods listed in Table I, then, go from the most searching, at the top, to the least searching, at the bottom.

There are two other data collection methods in common use which should be included here. These are the Method of Equal Appearing Intervals (17) and the Method of Successive Intervals. They may both be regarded as adaptations of the Method of Rank Order in which ties in rank are permitted. In the Method of Equal Appearing Intervals, the subject is instructed to place the stimuli in a given number of ranks, equally spaced psychologically, in the Method of Successive Intervals, he is merely given the number of ranks (piles) with no constraint on their 'spacing'. These methods although less searching than the Method of Rank Order, are not directly comparable to any of the other 'Order k ' methods. This might be more evident if one recognizes that 'Order k ' for $k < n-1$ yields a segment of one end of an I scale. The Method of Equal Appearing Intervals and the Method of Successive Intervals yield the entire I scale collapsed into steps which are or are not, respectively, assumed to be equal psychologically.

A more thorough development of the power structure on methods of collecting data, including methods not covered here, reveals that they are partially ordered with respect to the amount of information they contain about the behavior in question.

There are some general implications and interpretations which follow from the power structure on methods of collecting data considered from the general point of view of measurement theory. A selection of a method of collecting data and a method of analysis in any particular instance is dependent upon a resolution of the dilemma that the social scientist faces. Having resolved the dilemma to suit his purposes, he can select a method of collecting data and a method of analyzing it which will, for example, guarantee that he will end up with a unidimensional scale, or, if his objective is different, that will provide a test of dimensionality.

One of the basic issues in the interpretation of data is illustrated by the example of an individual who says, in three successive judgments,

TABLE I

Scale of Searchingsness for Methods of Collecting Data in Quadrant I

Method of collecting data	Number of genotypic classes of individuals in the case of one dimension	$\binom{n}{2} + 1$	Additional characteristics of the data	Methods of analysis
Method of Triads			In the case of one dimension these data provide more detailed information on the dispersion of an individual on the continuum and they contain the information as to whether an individual's judgments are consistent	The analysis of these data by the Unfolding Technique to obtain genotypic scales is dependent upon the generalization of the model to multidimensional unfolding, the analysis of these data to secure phenotypic scales is best done by application of Thurstone's Law of Comparative Judgment or one of its adaptations
Method of Paired Comparisons		$\binom{n}{2} + 1$	In the case of one dimension, these data provide information on the dispersion of an individual on the genotypic continuum. They contain the information as to whether the individual's judgments are transitive	
Method of Rank Order		$\binom{n}{2} + 1$	In the case of one dimension with adequate sampling these data provide information on metric relationships. In the case of 'Order 3,' the data contain information on the	The Unfolding Technique is appropriate for analyzing these data genotypically. The generalization of the Unfolding Technique to multidimensional unfolding has been accom
Order $n-1$ $n-2$		$\binom{n}{2}$		

plished only for certain special cases
The first published step in the gen
eralization is contained in Bennett
(1) Where the methods of analysis
appropriate to the Method of Suc
cessive Intervals, the Method of Equi
Appearing Intervals and the Method
of Rank Order are applied the result
is again a phenotypic scale of popu
larity

To analyze these data for genotypic
scale or latent attribute, the Parallel
ogram Method of analysis is appro
priate A method of matrix representa
tion and analysis suggested by
Festinger (11),† designed in the con
text of sociometric data is also ap
propriate here Analysis of these data
for phenotypic scales usually revolves
around counting the number of votes
and yields a 'popularity' scale

metric relationships of alternate pairs
of intervals between stimuli As k in
creases the metric information in
creases up to the Method of Rank
Order which yields the maximum
amount of metric information These
methods however containing so
much more additional information
are increasingly sensitive to both
multidimensionality and inadequate
sampling

In the case of one dimension with
adequate sampling these data will
yield a simple order of both stimuli
and people These methods differ also
in their sensitivity to multidimen
sionality and to unrepresentative
sampling A Pick k with k in the
immediate neighborhood of $n/2$ is
the most sensitive When unidimen
sionality is *not* present it is most
likely to be obtained in data with
little information therefore use an
 n small and k large

$$1 + nk - \frac{k(k+1)}{2}$$

k

•

•

•

•

3

3n5

Order 2

Pick 2

3

•

•

•

k

•

•

•

n^2

n^1

2 (n 1)

n^1

n^2

•

•

•

$n k + 1$

•

•

•

3

2

• I am indebted to Donald Meia Department of Mathematics University of Michigan for this general equation

† For additional developments see (9) (22) (23) and (24)

ments, that he prefers A to B , B to C , and C to A . One does not know whether the individual's judgments are inconsistent or whether they are intransitive. With a technique as searching as the Method of Triads, it appears possible to make a distinction between behavior which may be classified as error and behavior which requires explanation. By behavior which may be classified as error is meant behavior of a single individual which is random over replications of the stimulus situation. If the Method of Triads revealed that the individual was consistent and intransitive, it is incumbent upon the experimenter to accept this as experimental fact, regardless of the behavior of this individual on other stimuli or the behavior of other individuals on these stimuli. All conventional methods of measurement or scaling which classify intransitive judgments or other portions of data as error make one or more of the following assumptions: (1) that different individuals are replications of one another for the same stimulus situation, (2) that different stimuli are replications of one another for a given individual, (3) that a theory (level of measurement) is valid in spite of the data. It is now evident that these assumptions are neither necessary nor desirable unless the experimenter has resolved his dilemma by deciding to construct a unidimensional scale in spite of the data.

When a data collection method less searching than the Method of Triads is employed, a distinction between inconsistency and intransitivity in an individual's judgments is no longer possible unless one or more of the assumptions above are made. The Method of Paired Comparisons imposes consistency (reliability) on the judgments of an individual, and the Method of Rank Order further imposes that the paired comparisons be transitive. If the paired comparison judgments of an individual are transitive, the data may be expressed as a rank order with no loss of information. But if the data are collected by the Method of Rank Order, transitivity of the paired comparisons has been imposed on the behavior by the method of observing it and it is not known whether the behavior would have been transitive or not.

As we move down the searchingness scale, there is a series of successively decreasing numbers of elements in the rank order. Where the number of stimuli is n , the individual is asked to give his rank order of preference only for the first k ranks where $k < n$. Those

methods impose all the properties that the Method of Rank Order does but, because the rank order is incomplete, the missing segment is in effect determined for each individual by the judgments of the other individuals. In other words, since we have no information on the end segment of an individual's rank order, it immediately is compatible or in agreement with any obtained data. Thus, in all methods of collecting and analyzing data, that information in data *not collected* is always regarded as compatible with the information that was obtained.

If a simply ordered scale of stimuli and judges is desired, data with too much information in it may contain information incompatible with the desire. By the use of a method of collecting data which will provide less information, such as a 'Pick k ' method instead of an "Order k ," a simple order of all the stimuli may be constructed which is inferred to hold for all. This illustrates a general interpretation that may be given the searchingness structure on methods of collecting data. In one sense the searchingness structure may be regarded as a set of criteria for unidimensionality of behavior, the less searching being the weakest criteria and the most searching the most rigorous. Behavior which under a given criterion appears unidimensional will so appear for all weaker methods but may or may not satisfy unidimensionality under the stronger criteria.

There is another implication of this power structure on methods of collecting data—the fundamental principle that social science data are worth no more than the 'effort' expended by the judges in making their evaluations. This is illustrated throughout the power structure. It is easier for a judge, for example to "Pick 2" than to "Order 2." The latter contains the 'Pick 2' information and more. The principle is again illustrated by the relationship of the Method of Triads to the Method of Paired Comparisons.

The Method of Triads for relative behavior, task A , is at the top of the power structure for all the methods considered here and is minimal in the properties it imposes on the data. In fact this method of collecting data permits "almost anything to happen," and the inherent variability and other characteristics of the behavior are permitted to reveal themselves. The Method of Triads requires so much effort on the part of the judges, however, that it is impractical for a large number of stimuli, and this is probably the primary

reason for its being used so little. For the intensive investigation of Quadrant I behavior, over a moderate number of stimuli, it is the best of all methods presented here.

Because of the large amount of information in data collected by the Method of Triads, a judiciously selected portion of triads can be substituted for the Method of Paired Comparison when the latter method appears too formidable for the judges. This results from the fact that *each* triad may be converted into three paired comparisons with two degrees of freedom. One of the objections to the Method of Paired Comparisons is that it is tedious for an individual over a large number of stimuli. For $n = 20$, the number of paired comparisons required of an individual is 190. The full Method of Triads for 20 stimuli would require the individual to make judgments in 1140 triads, and each triad is the equivalent of three paired comparisons with two degrees of freedom. It is possible, however, to select 63 of these 1140 triads which, with one additional paired comparison, would be equivalent to the 190 paired comparisons, but the number of degrees of freedom would be 127 instead of 190. If the 190 degrees of freedom were required, it would take 95 triads which would be decomposable into 285 paired comparisons, so that some of the 190 different paired comparisons would be repeated.

This presentation of methods of collecting data has by no means exhausted the variety. Enough has been presented to permit the construction of a simple power structure, one of the directions for further generalization has been pointed out, some of the implications of the power structure have been indicated.

Quadrant IV

The presentation up to this point has been entirely in the context of Quadrant I, in which the individuals indicate comparative preferences between stimuli. These methods are, potentially at least, available for use in Quadrant IV also, with an appropriate change in instructions. In this quadrant the behavior of individuals consists of the comparative evaluation of stimuli with respect to an attribute. Thus, for 'Pick k ,' an individual would be asked to 'Pick the k most aggressive children in this group.'

The information in such judgments is not which of the stimuli is *nearer an ideal* of the judge on some underlying attribute, as in

Quadrant I, but which of the stimuli has *more* of some attribute. There is here, in principle, no assumed relationship between any ideal which the judge may or may not have and his task *B* judgment. Let this be made very clear. One individual may prefer candidate *A* to *B* and another one *B* to *A*, because each of these two judges has different ideals, for one judge, *A* is nearer his ideal, for the other, *B* is nearer. But if these two judges were asked whether *A* or *B* was more conservative, a better administrator or speaker, or more pro union, etc., it is not assumed that their ideals or the hypothetical preferences of the two judges on any of these attributes have any relationship to their judgments.

In concrete terms it is assumed that two individuals one in favor of and the other against universal military training will *not* for that reason, disagree as to which of the following statements is more favorable to universal military training (1) All men at the age of 18 should take one year of military training

(2) All men at the age of 18 should be urged to take one year of military training


The immediate objective of collecting data by the methods of Quadrant IV is to study the stimuli. Usually this means that the data are analyzed to obtain scale values for the stimuli on the attributes, commonly in the form of an interval scale. With such an objective most of the methods in the power structure would not be useful because the stimuli usually have sufficient spread on the relevant attribute that under Pick *k* the choices of most individuals would be the same and the stimuli *not* chosen could not be scaled. Hence the methods generally used are the more powerful methods which require each individual to make comparative judgments in many regions of the scale. These methods include the methods of Equal Appearing Intervals Successive Intervals Rank Order (34) Paired Comparisons (33) and Triads.

Data collected by these methods under task *B* instructions can not be analyzed by the Unfolding Technique because the data do not involve evaluation of stimuli with respect to an ideal. The method of analysis used on such data is usually the Law of Comparative Judgment (33) or an appropriate adaptation of it. One can of course use this latter method of analysis on paired comparison judgments collected under task *A*. In such a case the results of the analysis will bear a distinct algebraic relationship to the results

obtained from applying the Unfolding Technique to the same data. The Unfolding Technique yields a genotypic analysis of preferences; the Law of Comparative Judgment, or its modifications, yields a phenotypic analysis. These two analyses, then, have a predictable relation in this general theory. This is an example of a single set of data (collected by the Method of Paired Comparison, task *A*) which, depending upon the attitude or objective of the experimenter, may be located by his method of analysis in Quadrant I or Quadrant IV.

To extend the Unfolding Technique to the analysis of Quadrant IV data a somewhat different method of collecting data must be employed. This method may be called the Method of Similarities. In this method the individual is presented with the stimuli three at a time as in the Method of Triads. The instructions, however, are to judge which pair of the three is most alike and which pair is least alike. The instructions may or may not indicate an explicit attribute—*i.e.*, the instructions may be to judge which pair of a triad of cultures is more alike in their ethics, or the instructions may be merely to judge which pair is more alike. In either case the analysis of the data follows the same procedure. With the attribute explicit, however, the Method of Similarities will yield the latent structure underlying the stimuli for this attribute as perceived by an individual. With no explicit attribute, the method will yield the latent structure underlying the similarities and differences of stimuli as perceived by an individual.

This technique was used by Richardson as a method of collecting data in his two dimensional study of color (27). The method of analysis, however, was a carry over of Thurstone's Law of Comparative Judgment. The method suggested here for analyzing these data is quite different—namely, the Unfolding Technique. This difference in method of analysis is a direct consequence of the definition of the information contained in the data. To apply the Unfolding Technique to such data, the information they contain is obtained as follows. The responses of the individual to each such triad may be converted into three paired comparisons, each an element of a different I scale. For example, suppose an individual

responded to the triad  with the statements that *B* and *C*

were most alike and *B* and *D* least alike. If the individual had been asked whether stimulus *C* or *D* were more like *B*, the foregoing judgments imply that the individual would have said *C*. In effect, then, the individual was taking stimulus *B* as his ideal in the stimulus space and was saying that "from the point of view of *B*, *C* is preferred to *D*." This yields one of the paired comparison judgments that make up an *I* scale for the individual standing at the position of stimulus *B*.

In exactly the same manner his responses to this one triad provide one of the judgments for the individual's *I* scale when he is at stimulus *C* and one of the judgments for his *I* scale when he is at stimulus *D*. Thus, from the point of view of stimulus *C*, *B* is preferred to *D*, and from the point of view of *D*, *C* is preferred to *B*. From the individual's responses to the rest of the triads, an *I* scale, not necessarily transitive, may be constructed for the individual standing at each stimulus position.

Klingberg (19), in his study of the hostility relations among sovereign states, had each subject rank order ($n-1$) states in order of decreasing friendliness from the remaining state. Each state in turn was used as the standard. This technique necessarily yields transitive *I* scales, whereas the Method of Similarities may yield intransitive ones.

Data from either of the above methods may be analyzed by the Unfolding Technique to study the latent structure underlying the stimulus domain for a single individual. It is apparent, then, that such a method of collecting data in Quadrant IV contains information which permits it to be mapped into Quadrant I, and the methods of analysis appropriate to Quadrant I may be applied to these data. In the power structure of searchingness the Method of Similarities corresponds to the Method of Paired Comparisons and Klingberg's technique to the Method of Rank Order.

Quadrants II and III

In the power structure of methods of collecting data, "Pick 1" as a method was mentioned as a special case of the Method of Choice, corresponding to the method of Single Stimuli. This will now be discussed in more detail.

Consider an example of Quadrant III behavior, monotone

item a judge evaluating a subject by means of a rating scale. From the point of view of the information in the response, he may be said to be taking as his ideal his notion of the subject's ideal, and from that position he is picking one of the alternatives on the rating scale as being the stimulus nearest this ideal. When the ideal which the judge takes is actually his *own* ideal, task *A* explicitly, the behavior is classifiable as Quadrant IIa. But in terms of the information in each response, there is no difference between Quadrant II and Quadrant III, they are both "Pick 1" among a set of alternatives being evaluated with respect to an ideal.

Thus, monotone questionnaire or attitude items, typically of the Likert type (21) with alternatives running from "strongly agree" to "strongly disagree," are simply rating scales where the data are collected by "Pick 1." This type of data is tolerated even though it contains no information on the order of the alternatives because the experimenter has an *a priori* order which he regards as universally acceptable. This constitutes a generalization of the concept of "right answer." As far as the data are concerned there are nPn possible simple orders, as will be seen in the next section, but this does not matter to the experimenter because he knows which one is "right." If the data were collected by any of the other methods in the searchingness structure, tests for simple order of the alternatives and for common metric relations among the alternatives would be possible.

This way of looking at the Method of Single Stimuli is from the point of view of each item separately. Each item is regarded as representing an attribute, and its alternatives are the stimuli on this attribute among which the judge "picks 1." When there are only two alternatives to each item (endorse not endorse, pass fail, agree disagree, yes no, etc.), there is another way of looking at the Method of Single Stimuli—from the point of view of the group of items taken as a whole. In effect, the experimenter is regarding the items themselves, not their alternatives, as the stimuli, then the Method of Single Stimuli corresponds to "Pick any," with no constraint as in the Method of Choice.

There is no fundamental distinction between these two ways of looking at the Method of Single Stimuli. The fundamental distinctions that exist reside in the method of analysis selected by the experimenter and the implicit theory about the behavior that is

thereby implied. The methods of analysis designed for such data include the systems of Guttman (16), Lazarsfeld (20), and Test Theory (15). Their relations and distinctions will be brought out in the next section.

METHODS OF ANALYZING DATA

Paralleling the organization of the preceding section, we shall consider methods of analyzing data for the various methods of collecting data on the basis of their definition of the information contained in the data.

The methods of analyzing data appropriate to Quadrant I behavior, which include the Parallelogram Technique and the Unfolding Technique, will be discussed first; these will be followed by the methods appropriate to Quadrant IV behavior, which include the Law of Comparative Judgment and its derivatives and the extension of the Unfolding Technique to the Method of Similarities; finally the discussion will consider the methods appropriate to Quadrant II and III, which in the most general case is Lazarsfeld's system of theories, within which Guttman's theory and Test Theory are special cases.

Analysis of Quadrant I Data

THE PARALLELOGRAM TECHNIQUE. Parallelogram Analysis is specifically designed for analysis of data collected by one of the methods from "Pick $n-1$ " to "Pick 2." A matrix is constructed with items as columns and individuals as rows. If an individual endorses an item, an entry, X , is made in the cell at the intersection of the respective arrays. The rows and columns of this matrix are permuted in an attempt to collect the cells with entries in them in a diagonal band from top left to bottom right such that the entries in every row and in every column are adjacent. If this can be accomplished with an indecomposable matrix,¹⁰ the behavior of the individuals can be described by a simple order in which the order of the rows and

¹⁰ An indecomposable matrix has the property that there exists an arrangement of rows and columns such that the entries in every pair of successive rows have entries in at least one common column

the order of the columns represent the ordinal positions of individuals and items on a latent attribute. If such a pattern is not obtained, it may reflect error, multidimensionality, or unrepresentative sampling.

An illustration of this form of analysis is given in Figure 8 corresponding to the data collected by Pick 2 illustrated in Figure 2. The rows of Figure 8 each represent one individual from each class of phenotypic behavior.

A	B	C	D	E
X	X			
	X	X		
		X	X	
			X	X

FIG. 8. Parallelogram analysis of Pick 2 data.

A significant aspect of this type of analysis is that it is completely subordinate to the data. It does not map the data into a unidimensional continuum but, in effect, asks the data whether some of the conditions for a unidimensional continuum are satisfied. As a consequence, the technique of Parallelogram Analysis does not necessarily yield a simple order. It is a weak system of analysis in the sense discussed earlier.

THE UNFOLDING TECHNIQUE A change in the method of analysis occurs when we reach the Order 2 method of collecting data. Here the cell entry, previously an X, is either a 1 or a 2, to represent an individual's first or second choice. The analysis then consists of permuting rows and columns to form a parallelogram as in the analysis of Pick *k* data but in which the entries are digits as in Figure 9. The data analyzed in Figure 9 correspond to those illustrated in Figure 6.

The results of such an analysis are identical to those of Pick 2 except that with adequate sampling there are twice as many classes of individuals.

	A	B	C	D	E
1	1	2			
2	2	1			
		1	2		
		2	1		
			1	2	
			2	1	
				1	2
				2	1

FIG. 9. Unfolding analysis of "Order 2" data.

The permuting of rows and columns is continued until either a matrix is obtained which meets certain conditions or it is determined that no such matrix exists. If the conditions are met, then, as far as the power of that method of collecting data reveals; (1) the behavior may be regarded as a consequence of a single underlying latent attribute on which the stimuli have been at least simply ordered; (2) the judges have been placed in equivalence classes corresponding to their equivalent phenotypic behavior; (3) these classes have been ordered. Exactly what these conditions are will be made explicit after a discussion of the analysis of "Order 3" data.

Consider the classes of phenotypic behavior obtained from "Order 3" for five stimuli when there is a single common latent attribute underlying the phenotypic behavior, illustrated in Figure 7. For the analysis of such data, a matrix is constructed with cell entries of 1, 2, and 3 in each row indicating an individual's first, second, and third choice. The analysis then consists of permuting rows and columns to form a parallelogram, as in Figure 10.

The characteristics which such a matrix would possess, under the conditions of a single common latent attribute underlying the

choices of the individuals may now be made explicit for the general case of Order k "

- (1) The entries in each row (the integers from 1 to k) and in each column must be adjacent with no blanks
- (2) The entries in the first row and the first column must monotonically increase from left to right and from top to bottom respectively
- (3) The entries in the last row and the last column must monotonically decrease from left to right and from top to bottom respectively
- (4) The entries in all other columns must monotonically decrease and then increase from top to bottom

A	B	C	D	E
1	2	3		
2	1	3		
3	1	2		
3	2	1		
	2	1	3	
	3	1	2	
		1	2	3
		2	1	3
		3	1	2
		3	2	1

FIG 10 Unfolding analysis of Order 3 data

It is now evident in what sense the Parallelogram Technique for the analysis of Pick k is a special case of the Unfolding Technique for Order k . If x 's were substituted for the integers in Figure 10 the data would correspond to Pick 3 instead of Order 3 and discriminations among certain classes of individuals would vanish. In this instance the first four classes would collapse into

one class, the next two would collapse, and the last four classes would collapse, leaving only three classes of individuals, corresponding to Figure 4

This comparison reveals the difference in the searchingness of these methods of collecting data. If an individual under Order 3 had chosen *A*, *C*, and *B* as his first, second, and third choices the Unfolding Technique would reveal him as an aberrant individual in the sense that the latent attribute underlying his preferences is not the same as for the other individuals. Under Pick 3, this individual would be indistinguishable from the others in the first four classes that were collapsed together. The weaker condition for unidimensionality would be satisfied under Pick 3, but the stronger condition under 'Order 3' would not be satisfied. If such behavior were obtained, the social scientist would then be faced with the dilemma of whether to regard this individual as in error in his judgment or not. Data collected at the power level of Order 3 do not contain such information.

The analysis presented in Figure 10 has some further implications with respect to the genotypic structure underlying the preferences of individuals. These are in regard to the metric relations of the distances between stimuli on the *J* scale, as was pointed out in the preceding section.

This form of analysis, the Unfolding Technique, is applicable for all methods of collecting data in Quadrant I of the form

'Order k , $2 \leq k \leq n-1$ '. Ultimately for the Method of Rank Order corresponding to 'Order $n-1$ ', there is an integer in every cell of the matrix.

Throughout all of these methods the integrity of the data is maintained. No ordered metric or even simple order is necessarily obtained unless the data satisfy the required conditions. In the domain of social psychological variables the data will usually *not* satisfy these conditions and again the dilemma of the social scientist arises. Either he has to choose between his theory and his data if he believes or insists that an interval scale or a simple order holds or he has to use a method of collecting data which permits a distinction to be made between inconsistency and intransitivity in the judgments of each individual.

The Unfolding Technique is a method of discovering and

isolating a latent attribute underlying the preferences of a group of individuals or, in different terms is a method for discovering a genotypic scale. The latent attribute, if present, is represented by an ordered metric with both stimuli and individuals on the Joint scale. This is the only method available at present for analyzing choice behavior for latent attributes. The method is now being extended to unfolding in multidimensional space (1). When this is done, it will permit the analysis of preferences into two or more latent attributes—an analogue of multiple factor analysis.

The handling of data collected by the Method of Paired Comparisons is more involved unless the paired comparisons are transitive for each individual. In the latter case, the data may, of course, be converted to rank orders and analyzed as above. The theory for the analysis of intransitive paired comparisons has not yet been completed and there are several alternative psychological models all of which must be developed. Similarly the treatment of data collected by the Method of Triads, using no stronger psychological postulate than that given on page 497, is somewhat more involved because at this level of data collection the data contain information on random error within individuals. These further developments cannot profitably be pursued here.¹¹

The Unfolding Analysis of Quadrant I behavior collected by the Method of Successive Intervals or the Method of Equal Appearing Intervals requires a simple modification in procedure. In any row the digit representing the ordinal rank of a pile will be associated with as many columns as there were stimuli placed in that pile by that individual. The process of analysis is essentially similar to that given for the Order k methods of collecting data except that there is a certain relaxing of the conditions that the final matrix must satisfy. This is particularly true of the Method of Successive Intervals in which the piles have no *built in* metric relations.

The discussion of the methods of analyzing data collected by the Method of Choice task A has all been in terms of a weak system of analysis in which the data are regarded as paramount and a unidimensional analysis is obtained only if the data satisfy the necessary conditions. In general, as one goes down the scale of searchingness to methods of collecting data which are less searching

¹¹ For a more detailed treatment see Coombs (8 Chap. 7)

assumptions are substituted for data. These assumptions, in a general sense, are those necessary to make the data *not collected* compatible with the data collected.

It was mentioned earlier that the methods of collecting data in Quadrant I constituted criteria for unidimensionality of a latent attribute. It may now be seen more clearly in what sense this is so. The techniques from "Pick $n-1$ " up to and including "Order 2" are increasingly stringent criteria for a simple order. From "Order 3" up to and including "Order $n-1$," the Method of Rank Order, these criteria successively demand not only the same simple order but become more and more sensitive to metric relations. Thus, individuals may have the same simple order for a set of stimuli but different metric relations on them. Such individuals could not be placed on a common *continuum* without violating data.

The methods of analyzing the data of Quadrant I discussed up to this point are methods for discovering latent attributes (called genotypic scales) underlying preferences. These same data, however, may be used to construct *phenotypic* scales by any one of several systems of analysis developed by Thurstone. A phenotypic scale is a scale which best "represents" the data but does not "derive" them—that is, it does not in some sense go behind the data and draw inferences of a genotypic or explanatory nature but rather attempts to provide a simple *description* of all the data.

Methods of analysis for arriving at phenotypic scales for Quadrant I data include the systems of analysis specifically designed for each of the following methods of collecting data: the Method of Equal Appearing Intervals, the Method of Successive Intervals, the Method of Rank Order (34), and the Method of Paired Comparisons. The Method of Paired Comparisons is the most general of these methods, and Thurstone has developed the Law of Comparative Judgment for the construction of scales from such data.

These methods of analysis were designed by Thurstone for the scaling of stimuli with respect to an attribute, Quadrant IV behavior, and will be discussed below. They may, however, be applied to analyze Quadrant I behavior, appropriately collected. The relationship of such an analysis to an analysis of the *same data* by the Unfolding Technique will be pointed out below under the discussion of Group Scales.

Analysis of Quadrant IV Data

THURSTONE'S SCALING METHODS As previously indicated relative behavior involves a choice between stimuli. In the preceding subsection the choice involved reference to an "ideal" stimulus and hence the behavior reflects the relative preferences of an individual. In this section the behavior observed will be choice behavior but with reference to an attribute (Quadrant IV of Fig. 1). Hence the behavior reflects the individual's judgments on the relative magnitudes of two or more stimuli in some respect.

The immediate objective of collecting such data is to study the stimuli. Usually this involves using the data to determine the relative scale positions of the stimuli with respect to some attribute. The most usual methods of collecting data to achieve this purpose are those of Thurstone: the Method of Paired Comparisons, the Method of Rank Order, the Method of Successive Intervals, and the Method of Equal Appearing Intervals.

The procedures followed to construct a scale from the data collected by any of these methods are well known and available in the literature (14) and will not be repeated here. These are all strong systems which assume certain properties to hold for the information in the data, and the analysis yields an interval scale with the stimuli positioned on it. The usual application of interest to the social scientist, Case V of the Law of Comparative Judgment applied to data collected by the Method of Paired Comparisons, assumes that different individuals are replications of one another for the same stimulus situation. Stimuli must be used which are relatively indiscriminable so that there is some disagreement between judges. It is assumed that this disagreement is due to a given stimulus giving rise to a distribution of sensation magnitudes. Assuming that this distribution is normal and that the variability of the distribution of differences between pairs of such distributions is constant for all pairs (Case V) this latter variability is then used as the source for a unit of measurement and an interval scale may be built.

The methods of analysis designed for data collected by each of the other methods—the Method of Rank Order, the Method of Successive Intervals, and the Method of Equal Appearing Intervals—are essentially special cases of the Law of Comparative Judgment.

and are listed in the order in which a successively increasing number of assumptions are made or an increasing number of properties are imposed on the behavior by the method of observing it. These methods will always yield an interval scale unless the behavior has no error (i.e., all individuals agree on every judgment) or consists entirely of error (i.e., individuals split 50-50 on every judgment). It is of interest to note in passing that the methods of Thurstone require relatively indiscriminable stimuli whereas the Unfolding Technique is much more suitable for completely discriminable than for relatively indiscriminable stimuli.

These methods of Thurstone are commonly employed to construct an attitude scale with statements of opinion ranging from pro to anti. The scale so constructed has the stimuli located on it but not the judges. One then assumes that the scale obtained holds for all the judges or for a different group of individuals and it is readministered under a method appropriate to Quadrant I or II to locate the individuals on the scale. Two experimental operations for collecting data are required to yield a joint distribution, and the stimuli must be relatively indiscriminable in order that there be some error variance to yield a unit of measurement.

No better methods have been devised for constructing an interval scale for measuring attitudes. The Unfolding Technique applied to such behavior may yield the joint distribution in one experimental operation but it will be at best an ordered metric and not an interval scale. Furthermore, the present writer's experience has shown that it is much more likely to imply that no such yardstick exists. If, for reasons of belief or convenience, one requires that a social psychological attribute be measured on a straight line by use of the real numbers, it appears to be necessary to use techniques of observation and analysis which embody sufficient assumptions and classify sufficient data as error to ensure such a result.

THE METHOD OF SIMILARITIES In the preceding section, on methods of collecting data, the Method of Similarities was presented as a method by which the Unfolding Technique could be extended to Quadrant IV behavior. On the basis of the information contained in such data, described in the preceding section, the judgments of the individual may be converted into paired comparison judgments with respect to an ideal in which the individual is regarded as taking each stimulus in turn as his ideal. If the paired comparison

judgments are transitive for each ideal, then a rank order I scale is obtained. Such data may be analyzed as described for 'Order k ' in which $l = n - 1$, each row of the matrix representing the rank order from a given ideal. The entire matrix then represents the behavior of a single individual over the entire stimulus space. This technique permits a relatively rigorous and intensive study of a single individual.

THE GROUP SCALE It has been pointed out several times that data collected by certain methods in Quadrant I (e.g., the Method of Paired Comparisons) may be analyzed by the Unfolding Technique or by the Law of Comparative Judgment. The relationship between these forms of analysis will now be shown.

If the Method of Paired Comparisons is used to collect data in Quadrant I, the judgments of individuals represent preferences on each pair of stimuli. If the Law of Comparative Judgment is used to analyze such paired comparison data, the result is a scale which in some statistical sense (25) is most descriptive of the preferences of the individuals taken as a group. The application of such a technique involves regarding preference as an attribute of stimuli (e.g., preferability or popularity) and, in effect, the experimenter is regarding such data as evaluation of stimuli with respect to an attribute and placing it in Quadrant IV instead of Quadrant I.

Application of the Unfolding Technique to the same data may yield, if the appropriate conditions are satisfied, a Joint scale on which the stimuli and the individual judges are located instead of just the stimuli. The unfolded J scale is a genotypic scale representing an inferred latent attribute underlying the preferences of the individuals. The Law of Comparative Judgment solution is a phenotypic scale *descriptive* of the preferences.

In the more formal development of this general scaling theory, the individual's I scale, which is observed at best as a rank order, may hypothetically be regarded as derived from a ratio scale of 'preferability' for each individual. Let us define a Group scale as a mean of all the I scales. In the special case in which the I scales arise from a common Joint scale, a Group scale is a Joint scale folded in the middle. It can then be shown that a Law of Comparative Judgment solution to preference data represents an approximation to such a Group scale. These theoretical relations have been tested in several experiments and have been borne out. Thus, in

the case in which a single latent attribute underlies the phenotypic preferences of individuals, the Unfolding Technique yields a Joint scale with individuals and stimuli located on it the Law of Comparative Judgment solution represents the same scale, folded approximately in the middle, with only the stimuli remaining on it

This discussion of Group scales has had explicit reference to preferential judgments collected by the Method of Paired Comparisons and analyzed by the Law of Comparative Judgment as well as by the Unfolding Technique It should be apparent that classifying data collected by the Method of Rank Order, the Method of Successive Intervals, and the Method of Equal Appearing Intervals in Quadrant IV and analyzing such data by the appropriate methods simply represent different approximations to the Group scale Data collected by any of these latter methods may also be analyzed by the corresponding case of the Unfolding Technique and a unidimensional Joint scale obtained under the appropriate conditions

The theory and the computational analysis of data collected by the Method of Paired Comparisons the Method of Rank Order, the Method of Successive Intervals, or the Method of Equal Appearing Intervals and analyzed by their appropriate Quadrant IV method of analysis make no distinction between task *A* and task *B* For these methods, all behavior is task *B* Only certain of these data, however, may be classified in Quadrant I and unfolded, and that is when these methods are used to observe task *A* behavior, the evaluation of stimuli with respect to an ideal

Analysis of Quadrant II and III Data

As was pointed out in the preceding section, the data associated with these two quadrants are those collected by the Method of Single Stimuli It was further pointed out that such data (the experimentally independent successive responses to a number of items by an individual choosing one alternative as his response to each item) constituted a special case, "Pick 1," of the Method of Choice

From the point of view of the information contained in such data, two types of items were distinguished monotone items, associated with Quadrant IIa and III, and nonmonotone items, associated with Quadrant IIb The information contained in monotone items, in the particular case of two alternatives, is whether the ideal

of the individual is greater or less than the position of the item on the genotypic scale

In the case of the nonmonotone items, the information contained in the data is which of the alternatives to an item is nearer the individual's ideal in *any* direction, rather than on one side of him as in the case of monotone items. Consequently, if the endorse alternative is 'too far away' in any direction, the phenotypic behavior is "not endorse, and hence there may be several distinct kinds of genotypic individuals responding identically phenotypically.

The kinds of data usually classified in Quadrants IIa or III include mental test data, attitude or questionnaire items with Likert type alternatives, items of a cumulative¹² nature, and rating scale data. The kinds of data usually classified in Quadrant IIb include statements of opinion ranging from pro to anti, administered by the Method of Single Stimuli with the alternatives "agree" or "not agree."

Lazarsfeld is in the process of constructing two related systems for the analysis of Method of Single Stimuli data. These systems are his Latent Distance Model and Latent Structure Analysis.¹³ These systems are complex and as yet relatively undeveloped so in many instances they are not practicable. Although these systems are not computationally feasible this constitutes only a transient difficulty. Lazarsfeld's system is actually a theory of theories, a metatheory, of methods of analyzing irrelative behavior. Lazarsfeld's system is of such generality that it provides a theoretical framework within which specific methods of analyzing data collected by the Method of Single Stimuli may be understood as special cases. Both Guttman's scaling theory and mental test theory will be presented in this context.

Viewing various methods of analysis as special cases of a more general theory reveals the fact that an experimenter in selecting a method of analysis is selecting a theory about behavior. The data are either *asked* to satisfy or *forced* to satisfy the theory, depending on the strength or the postulational basis of the specific theory.

¹² See for example Stouffer (31 p. 141).

¹³ This distinction is an arbitrary one based on the use of discontinuous trace lines in the Latent Distance Model and continuous trace lines in Latent Structure Analysis. The distinction is convenient here because Guttman's scaling theory is a special case of the first and mental test theory a special case of the second.

underlying the method of analysis selected. The common immediate objective of these techniques is to convert the information in the data to positions on an ordinal or interval scale. The techniques of analysis differ only in the properties or constraints they impose on the information in the data, the techniques range from very weak systems, which make the least assumptions and may not even yield a simply ordered scale, to very strong systems, which make sufficient assumptions to guarantee an interval scale.

LAZARSFELD'S LATENT DISTANCE MODEL This model is specifically designed for the analysis of Quadrant IIa and Quadrant III data. In the simplest form of this model, the attribute continuum is assumed to be dichotomized at some point by an item such that all individuals on one side of that point have a probability p_j of endorsing or passing the item j and all individuals on the other side

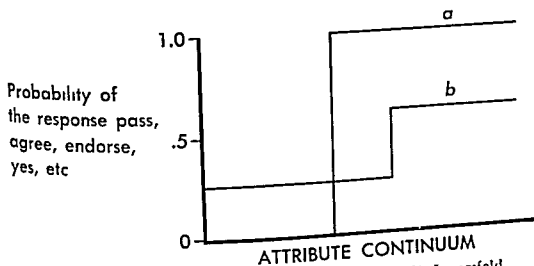


FIG. 11. Example of (a) Guttman type item and (b) Lazarsfeld type item

of that point have the probability $1 - p_j$ of endorsing or passing the item (cf Fig. 11). This model is readily generalizable to more than two latent classes for other than dichotomous items.

It is immediately apparent that Guttman's scalogram technique is a special case of the dichotomous item. The Guttman model requires that p_j be zero or one (cf Fig. 11).

In different terms, Guttman's system of analysis requires perfectly reliable items (perfect internal consistency), whereas Lazarsfeld can handle less than perfectly reliable items. By permitting

of the individual is greater or less than the position of the item on the genotypic scale

In the case of the nonmonotone items, the information contained in the data is which of the alternatives to an item is nearest the individual's ideal in *any* direction, rather than on one side of him, as in the case of monotone items. Consequently, if the 'endorse' alternative is "too far away" in any direction, the phenotypic behavior is "not endorse," and hence there may be several distinct kinds of genotypic individuals responding identically phenotypically.

The kinds of data usually classified in Quadrants IIa or III include mental test data, attitude or questionnaire items with Likert-type alternatives, items of a 'cumulative'¹² nature, and rating scale data. The kinds of data usually classified in Quadrant IIb include statements of opinion ranging from pro to anti, administered by the Method of Single Stimuli with the alternatives "agree" or "not agree."

Lazarsfeld is in the process of constructing two related systems for the analysis of Method of Single Stimuli data. These systems are his Latent Distance Model and Latent Structure Analysis.¹³ These systems are complex and as yet relatively undeveloped so in many instances they are not practicable. Although these systems are not computationally feasible, this constitutes only a transient difficulty. Lazarsfeld's system is actually a theory of theories, a metatheory, of methods of analyzing irrelative behavior. Lazarsfeld's system is of such generality that it provides a theoretical framework within which specific methods of analyzing data collected by the Method of Single Stimuli may be understood as special cases. Both Guttman's scaling theory and mental test theory will be presented in this context.

Viewing various methods of analysis as special cases of a more general theory reveals the fact that an experimenter in selecting a method of analysis is selecting a theory about behavior. The data are either *asked* to satisfy or *forced* to satisfy the theory, depending on the strength or the postulational basis of the specific theory.

¹² See for example Stouffer (31 p. 141).

¹³ This distinction is an arbitrary one based on the use of discontinuous trace lines in the Latent Distance Model and continuous trace lines in Latent Structure Analysis. The distinction is convenient here because Guttman's scaling theory is a special case of the first and mental test theory a special case of the second.

error or inconsistency in behavior, Lazarsfeld's system can yield a solution when Guttman's cannot. Lazarsfeld's solution provides a set of at least two latent classes on the genotypic level and a probability that each response pattern is associated with each latent class. When the data satisfy the conditions for a simply ordered scale, Lazarsfeld's system reduces to Guttman's in that the probability of a specific response pattern's being associated with a specific segment of the hypothetical continuum is *zero or one*.

The Guttman scalogram technique is used to analyze Quadrant Ila data to determine whether the conditions for a simply ordered scale are satisfied. The method is closely related to the Parallelogram Technique in that a matrix is constructed in exactly the same manner and the rows and the columns are permuted. But inasmuch as these items are monotone items instead of nonmonotone, the final matrix, if ordinality is satisfied, has all the X's on one side of the diagonal (including the diagonal) and all the blank cells on the other side of the diagonal. For this reason this method of analysis may be called *Triangular Analysis*.

Guttman's scaling theory constitutes a strict adherence to the logical structure of an ordinal scale. If all the data do not satisfy the conditions, he calls the result either a quasi scale or nonscale type. The latter is really a partial order. If an ordinal scale is insisted upon, part of the data must be rejected—either individuals, stimuli, or both. Otherwise the latent distance model, which can classify some of the data as error and still yield an ordinal scale, must be used.

The result of a Guttman analysis on data which satisfy the necessary conditions is an ordinal scale with the stimuli and the response patterns of the individuals simply ordered. The technique as conventionally used is applied to dichotomous items. If the items contain more alternatives than two, the surest way to find an ordinal scale is to group the alternatives to form that dichotomy which best happens to satisfy the conditions for an ordinal scale. This procedure takes advantage of every error, random or biased, in the data but has commonly been followed because of the generally unsatisfactory results of trying to scale all the alternatives.

It is possible to show that the scaling of the alternatives for different items by Guttman's technique requires an additional condition above the ordinal property of the scale. It is not feasible

to develop this here, but the general principle is that it is not the alternatives which should be scaled but the section line between two adjacent alternatives of an item. The alternatives may themselves appear to be scaled only if the alternatives of each item mesh, like the teeth in gears, with the alternatives of the other items. This condition can be called "orderly interlocking." If this condition does not hold, the conditions for an ordinal scale may exist but no ordinal scale will be found in the data unless the boundaries of the alternatives are scaled instead of the alternatives themselves. Thus, an item with *five* alternatives has *four* scale positions, one between alternative *a* and *b*, one between *b* and *c*, *c* and *d*, and *d* and *e*. In this manner it is also possible to handle, in one scale analysis, questionnaire items which have varying numbers of alternatives.

It has previously been pointed out that every item in the Method of Single Stimuli, task *A*, monotone stimuli is really a disguised rating scale. The experimenter must select an *a priori* order among the alternatives to the item. The individual in responding, simply informs the experimenter of his location on the rating scale. Methods of analyzing Quadrant IIa data are consequently simply methods of analyzing the task *A* responses of a number of individuals to a number of rating scales. Hence, from the abstract point of view of measurement theory, rating scale data can be mapped into Quadrant IIa and hence the methods of Guttman and Lazarsfeld are applicable to the analysis of rating scale data.

There is one characteristic of Guttman's scaling technique which is either an advantage or a disadvantage depending on one's point of view. If the data do not meet the necessary conditions for an ordinal scale, one will not be obtained. This is a good characteristic if one is interested in the study of behavior of individuals and the nature of attributes, for the technique does not force a more powerful system on the data than the data satisfy. On the other hand, if an ordinal scale is demanded, this characteristic is a disadvantage. In the latter case, if an ordinal scale is not found, the alternatives are to reject individuals or stimuli or change the responses of some individuals to some items—i.e., classify them as error.

Guttman, in his general theory, makes much of a theory of components which are successive sources of the variance of individuals' behavior. He particularly makes much of the psychological interpre-

error or inconsistency in behavior, Lazarsfeld's system can yield a solution when Guttman's cannot. Lazarsfeld's solution provides a set of at least two latent classes on the genotypic level and a probability that each response pattern is associated with each latent class. When the data satisfy the conditions for a simply ordered scale, Lazarsfeld's system reduces to Guttman's in that the probability of a specific response pattern's being associated with a specific segment of the hypothetical continuum is *zero* or *one*.

The Guttman scalogram technique is used to analyze Quadrant Ila data to determine whether the conditions for a simply ordered scale are satisfied. The method is closely related to the Parallelogram Technique in that a matrix is constructed in exactly the same manner and the rows and the columns are permuted. But inasmuch as these items are monotone items instead of nonmonotone, the final matrix, if ordinality is satisfied, has all the X's on one side of the diagonal (including the diagonal) and all the blank cells on the other side of the diagonal. For this reason this method of analysis may be called Triangular Analysis.

Guttman's scaling theory constitutes a strict adherence to the logical structure of an ordinal scale. If all the data do not satisfy the conditions, he calls the result either a quasi scale or nonscale type. The latter is really a partial order. If an ordinal scale is insisted upon, part of the data must be rejected—either individuals, stimuli, or both. Otherwise the latent distance model, which can classify some of the data as error and still yield an ordinal scale, must be used.

The result of a Guttman analysis on data which satisfy the necessary conditions is an ordinal scale with the stimuli and the response patterns of the individuals simply ordered. The technique as conventionally used is applied to dichotomous items. If the items contain more alternatives than two, the surest way to find an ordinal scale is to group the alternatives to form that dichotomy which best happens to satisfy the conditions for an ordinal scale. This procedure takes advantage of every error, random or biased, in the data but has commonly been followed because of the generally unsatisfactory results of trying to scale all the alternatives.

It is possible to show that the scaling of the alternatives for different items by Guttman's technique requires an additional condition above the ordinal property of the scale. It is not feasible

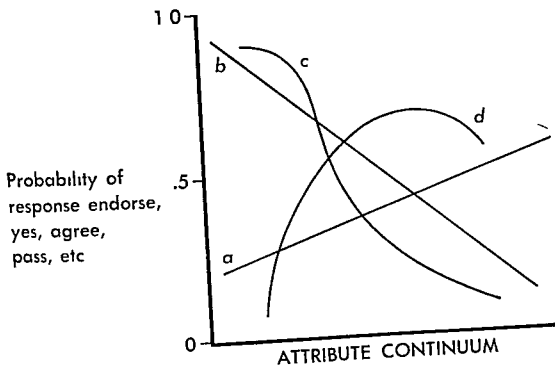


FIG 12 Illustration of monotone (*a b, c*) trace lines and non monotone (*d*) trace lines in Lazarsfeld's Latent Structure Theory

monotone items by the use of a trace line which is a continuous monotonic function of a latent attribute continuum—e.g., a straight line

This concept of a trace line is applicable to nonmonotone items by use of polynomial trace functions (13). The trace function may be assumed to be (*a*) a polynomial in several variables, (*b*) a linear combination of several variables, or (*c*) a polynomial in a single variable. Thus an item in which two extreme groups on a single attitude variable respond alike and the intermediate group differently would correspond to a parabolic trace line. This generalization permits any of an infinite variety of trace lines to be assumed. By virtue of a separate trace associated with each item, it is possible to apply Latent Structure Analysis, at least in principle, to data obtained by the Method of Single Stimuli from a mixture of monotone and nonmonotone items. In mental testing, the items are dichotomous—pass fail—and the patterns of response are mapped into the integers by counting all the favorable (passing) elements in a response pattern. Thus a potentially great variety of different response patterns are all mapped on the same integer. For example,

tations which he gives to the first and second components. The first component is the simply ordered scale of stimuli and individuals. The second component is a U shaped function which he interprets as 'intensity'. He suggests that individuals at the extremes of the attitude continuum feel more strongly about their attitude than those in the middle region, in the neighborhood of indifference. This was an early suggestion of Katz (18) and is a reasonable psychological hypothesis. But the interpretation of the second component as a *measure* of this intensity of feeling must be very carefully validated experimentally. As indicated by the class of data to which Guttman's theory applies, mathematical analysis into components is as valid for mental test behavior as for attitude data collected on monotone items. These same mathematical components would be found in the behavior of individuals on an arithmetic test, for example, but the same psychological interpretation of intensity of feeling about one's arithmetic ability would not necessarily follow. The interpretation to be given components may be different for different domains of behavior.

LAZARSFELD'S LATENT STRUCTURE ANALYSIS Instead of regarding the underlying attribute as having discrete steps or classes, in Lazarsfeld's Latent Structure Analysis (20), the underlying theory is that of a continuous gradation in the latent attributes. The analysis, as in the latent distance model, yields discrete latent classes, but the basic psychological theory requires an underlying continuum. Latent structure analysis is Lazarsfeld's general system for the analysis of non monotone items, Quadrant IIb data, but in which monotone items are a special case. It is in this sense that Lazarsfeld's general theory is a theory of theories for irrelative, task *A*, behavior. In principle his theory is applicable not only to monotone and nonmonotone items but also to combinations of them in the same test or questionnaire.

Latent Structure Analysis has the basic concept of a trace line associated with each item which represents the probability of an individual at any point on the latent attitude continuum responding favorably to the given item (Fig. 12).

In the early stages of the development of the theory, the selection of a function (straight lines, parabola, etc.) for the trace line was an *a priori* decision, but there is an expectation that this will become objective and unique (12). This system is applicable to

tations which he gives to the first and second components. The first component is the simply ordered scale of stimuli and individuals. The second component is a U shaped function which he interprets as intensity. He suggests that individuals at the extremes of the attitude continuum feel more strongly about their attitude than those in the middle region in the neighborhood of indifference. This was an early suggestion of Katz (18) and is a reasonable psychological hypothesis. But the interpretation of the second component as a *measure* of this intensity of feeling must be very carefully validated experimentally. As indicated by the class of data to which Guttman's theory applies, mathematical analysis into components is as valid for mental test behavior as for attitude data collected on monotone items. These same mathematical components would be found in the behavior of individuals on an arithmetic test for example, but the same psychological interpretation of intensity of feeling about one's arithmetic ability would not necessarily follow. The interpretation to be given components may be different for different domains of behavior.

LAZARSFELD'S LATENT STRUCTURE ANALYSIS Instead of regarding the underlying attribute as having discrete steps or classes, in Lazarsfeld's Latent Structure Analysis (20) the underlying theory is that of a continuous gradation in the latent attributes. The analysis, as in the latent distance model, yields discrete latent classes, but the basic psychological theory requires an underlying continuum. Latent structure analysis is Lazarsfeld's general system for the analysis of non-monotone items. Quadrant IIb data, but in which monotone items are a special case. It is in this sense that Lazarsfeld's general theory is a theory of theories for irrelative task *A* behavior. In principle his theory is applicable not only to monotone and nonmonotone items but also to combinations of them in the same test or questionnaire.

Latent Structure Analysis has the basic concept of a trace line associated with each item which represents the probability of an individual at any point on the latent attitude continuum responding favorably to the given item (Fig. 12).

In the early stages of the development of the theory, the selection of a function (straight lines, parabola, etc.) for the trace line was an *a priori* decision, but there is an expectation that this will become objective and unique (12). This system is applicable to

mental test except that in the latter the response categories, *fail* and *pass*, are mapped into 0 and 1, respectively

Likert discusses a more "refined" scoring procedure called the sigma method of scoring. The assumption is that the distribution on the underlying attribute is normal. The percentage of individuals in a given category is converted into a sigma value and that value is used instead of the integers in the simplified scoring system. Mapping the response categories into the real numbers like this instead of into the integers has been tried in mental test techniques also but in this area has not proved worthwhile, particularly in view of the labor involved compared with the simple weights.

This problem of arriving at scale scores on an interval scale has conventionally been treated primarily as an empirical problem and by means of statistical criteria—e.g., reliability, validity, etc. As indicated earlier, the level of measurement into which the data are cast is an intrinsic part of any theory about the data. If a given set of data should actually satisfy only a partial order, it can always be mapped into an interval scale by defining a constant and common unit of measurement. Such a scale will usually exhibit a significant degree of reliability in the sense of a sufficiently high ratio of the variability between individual's scale positions to variability within. In such a circumstance, statistically significant relations can be obtained with other similarly defined variables and hence these procedures have tremendously practical value for applied problems and lead to an effective actuarial science.

Lazarsfeld's system is also a generalization of Thurstone's multiple factor analysis (35) in that it is not confined to unidimensional analyses and permits other assumptions to be made than that of linear summation, which underlies factor analysis. Lazarsfeld makes use of a generalized concept of partial correlation between items whereas multiple factor analysis makes use of only the zero order correlations. It is this element of Lazarsfeld's system which multiplies the computational complexity because the system, in principle, may make use of partials of high order.

In this discussion of Quadrants II and III, it has been pointed out that Lazarsfeld's general theory is actually a variety of theories about irrelative behavior and that specific systems of analysis designed ad hoc for particular objectives are special cases of this more

a score of *two* could be obtained by passing any two items *PPFF*, *FPFF*, *FPFI*, *ITPP*, etc

If the conditions for a Guttman scale hold, a one to one correspondence exists between a response pattern and an integer corresponding to an ordinal position on a unidimensional continuum. To convert such a scale into an interval scale, an assumption must be made such as that the increments in ability between the ordinal positions of adjacent items on the scale are equal.

When conditions for a Guttman scale are not met, there is more than one response pattern mapped onto the same integer and the scale is represented by a partial order, what Guttman calls a non scale type. The rationale for mapping such a partial order on an interval scale is somewhat more involved, since it is based on a concept of error and random fluctuations.

The concept of error commonly used in test theory corresponds to the assumption that the trace line of a mental test item is the integral of a normal curve (*cf.* trace line *c* in Figure 12), although Carroll (5) is currently working on the theory of a linear trace line for mental tests. If the items were perfectly reliable, the variance of the normal curve would be zero and the trace line would take the form of that shown for item *a* in Figure 11. The rationale underlying test theory is critically evaluated by Reese (26) and Thomas (32).

Likert (21) has suggested a technique for arriving at a score on an attitude questionnaire which is a simple extension of mental testing technique. Likert's system is necessarily confined to monotone items but is an extension of the scoring methods of mental testing in that degrees of "endorsement" are obtained in the data. The technique is to take a statement of opinion sufficiently extreme so that it cannot act nonmonotonically—that is, so that there are not likely to be any people so extreme that they would reject the item for not being sufficiently extreme. To such an item, a degree of endorsement is obtained, typically from *strongly agree* to *strongly disagree*. The simplified scoring method is to map these alternatives into the integers—e.g., 1 to 5—so that 1 represents the extreme pro answer and 5 the extreme anti. The individual's score, then, may be simply the sum of those integers into which his responses have been mapped. Obviously this corresponds identically to scoring a

mental test except that in the latter the response categories, *fail* and *pass*, are mapped into 0 and 1, respectively

Likert discusses a more 'refined' scoring procedure called the sigma method of scoring. The assumption is that the distribution on the underlying attribute is normal. The percentage of individuals in a given category is converted into a sigma value and that value is used instead of the integers in the simplified scoring system. Mapping the response categories into the real numbers like this instead of into the integers has been tried in mental test techniques also but in this area has not proved worthwhile, particularly in view of the labor involved compared with the simple weights.

This problem of arriving at scale scores on an interval scale has conventionally been treated primarily as an empirical problem and by means of statistical criteria—e.g., reliability, validity, etc. As indicated earlier, the level of measurement into which the data are cast is an intrinsic part of any theory about the data. If a given set of data should actually satisfy only a partial order, it can always be mapped into an interval scale by defining a constant and common unit of measurement. Such a scale will usually exhibit a significant degree of reliability in the sense of a sufficiently high ratio of the variability between individual's scale positions to variability within. In such a circumstance, statistically significant relations can be obtained with other similarly defined variables and hence these procedures have tremendously practical value for applied problems and lead to an effective actuarial science.

Lazarsfeld's system is also a generalization of Thurstone's multiple factor analysis (35) in that it is not confined to unidimensional analyses and permits other assumptions to be made than that of linear summation, which underlies factor analysis. Lazarsfeld makes use of a generalized concept of partial correlation between items whereas multiple factor analysis makes use of only the zero order correlations. It is this element of Lazarsfeld's system which multiplies the computational complexity because the system, in principle, may make use of partials of high order.

In this discussion of Quadrants II and III, it has been pointed out that Lazarsfeld's general theory is actually a variety of theories about irrelative behavior and that specific systems of analysis designed ad hoc for particular objectives are special cases of this more

a score of *two* could be obtained by passing any two items *PPFF*, *FPFF*, *FPFP*, *FFPP*, etc

If the conditions for a Guttman scale hold, a one to-one correspondence exists between a response pattern and an integer corresponding to an ordinal position on a unidimensional continuum. To convert such a scale into an interval scale, an assumption must be made such as that the increments in ability between the ordinal positions of adjacent items on the scale are equal.

When conditions for a Guttman scale are not met, there is more than one response pattern mapped onto the same integer and the scale is represented by a partial order, what Guttman calls a non scale type. The rationale for mapping such a partial order on an interval scale is somewhat more involved, since it is based on a concept of error and random fluctuations.

The concept of error commonly used in test theory corresponds to the assumption that the trace line of a mental test item is the integral of a normal curve (*cf.* trace line *c* in Figure 12), although Carroll (5) is currently working on the theory of a linear trace line for mental tests. If the items were perfectly reliable, the variance of the normal curve would be zero and the trace line would take the form of that shown for item *a* in Figure 11. The rationale underlying test theory is critically evaluated by Reese (26) and Thomas (32).

Likert (21) has suggested a technique for arriving at a score on an attitude questionnaire which is a simple extension of mental testing technique. Likert's system is necessarily confined to monotone items but is an extension of the scoring methods of mental testing in that degrees of 'endorsement' are obtained in the data. The technique is to take a statement of opinion sufficiently extreme so that it cannot act nonmonotonically—that is so that there are not likely to be any people so extreme that they would reject the item for not being sufficiently extreme. To such an item, a degree of endorsement is obtained, typically from *strongly agree* to *strongly disagree*. The simplified scoring method is to map these alternatives into the integers—e.g., 1 to 5—so that 1 represents the extreme pro answer and 5 the extreme anti. The individual's score, then, may be simply the sum of those integers into which his responses have been mapped. Obviously this corresponds identically to scoring a

mental test except that in the latter the response categories, *fail* and *pass*, are mapped into 0 and 1, respectively

Likert discusses a more refined scoring procedure called the sigma method of scoring. The assumption is that the distribution on the underlying attribute is normal. The percentage of individuals in a given category is converted into a sigma value and that value is used instead of the integers in the simplified scoring system. Mapping the response categories into the real numbers like this instead of into the integers has been tried in mental test techniques also but in this area has not proved worthwhile, particularly in view of the labor involved compared with the simple weights.

This problem of arriving at scale scores on an interval scale has conventionally been treated primarily as an empirical problem and by means of statistical criteria—e.g. reliability, validity, etc. As indicated earlier, the level of measurement into which the data are cast is an intrinsic part of any theory about the data. If a given set of data should actually satisfy only a partial order, it can always be mapped into an interval scale by defining a constant and common unit of measurement. Such a scale will usually exhibit a significant degree of reliability in the sense of a sufficiently high ratio of the variability between individual's scale positions to variability within. In such a circumstance statistically significant relations can be obtained with other similarly defined variables and hence these procedures have tremendously practical value for applied problems and lead to an effective actuarial science.

Lazarsfeld's system is also a generalization of Thurstone's multiple factor analysis (35) in that it is not confined to unidimensional analyses and permits other assumptions to be made than that of linear summation which underlies factor analysis. Lazarsfeld makes use of a generalized concept of partial correlation between items whereas multiple factor analysis makes use of only the zero order correlations. It is this element of Lazarsfeld's system which multiplies the computational complexity because the system in principle may make use of partials of high order.

In this discussion of Quadrants II and III, it has been pointed out that Lazarsfeld's general theory is actually a variety of theories about irrelative behavior and that specific systems of analysis designed ad hoc for particular objectives are special cases of this more

prefer as friends, with no restriction on the number an individual picks, the behavior may be irrelative, task A , nonmonotone. If a restriction is placed on the number to be chosen, the behavior becomes relative (Method of Choice), task A , and the methods of analysis appropriate to Quadrant I are applicable.

BIBLIOGRAPHY

- 1 Bennett J F *A method for determining the dimensionality of a set of rank orders* Ph D thesis Univ of Michigan 1951
- 2 Bridgman P W *The logic of modern physics* New York Macmillan 1927
- 3 Bronfenbrenner U A constant frame of reference for sociometric research Part I *Sociometry*, 1943 6 363-397
- 4 Carnap R *Foundations of logic and mathematics* Chicago Univ of Chicago Press 1939
- 5 Carroll J B Problems in the factor analysis of tests of varying difficulty *Amer Psychologist*, 1950 5, 369 Abstract
- 6 Coombs C H Psychological scaling without a unit of measurement *Psychol Rev* 1950 57, 145-158
- 7 ——— Mathematical models in psychological scaling *J Amer Stat Assoc*, 1951 46, 480-489
- 8 ——— A theory of psychological scaling *Univ of Michigan Engineering Research Inst Bull No 34* Ann Arbor Univ of Michigan Press 1952
- 9 Criswell J H The measurement of group integration *Sociometry* 1947 10 259-267
- 10 Dushnik B Concerning a certain set of arrangements *Proc of the Amer Math Society*, 1950 1, 788-796
- 11 Festinger L The analysis of sociograms using matrix algebra *Hum Relat*, 1949 2, 153-158
- 12 Green B F A general solution for the latent class model of latent structure analysis *Psychometrika*, 1951 16, 151-166

general theory The advantage of an abstract approach to methods of collecting and methods of analyzing data, from the point of view of measurement theory, lies in its permitting the comparison of different techniques and their interrelations, it also permits generalizing their applicability

For example, from this abstract point of view, the data from certain learning experiments can be classified as Quadrant IIa data A conditioning experiment in the abstract corresponds to a mental test given backwards The conditioned stimulus and the unconditioned stimulus together may be regarded as the stimulus situation If the response occurs after the conditioned stimulus and before the unconditioned stimulus, the item is 'passed', if not, it is 'failed' The stimuli are presented in the order of the most "difficult" item first and, as conditioning progresses, the items become 'easier' The analogy is now clear If one wants to use the strong system of test theory, each individual's performance is mapped into the integers by counting trials to learn This requires the assumption that the increment in learning between all pairs of successive trials is constant within and among individuals

In conditioning experiments, the presentation of stimuli may be continued until the individual gets a certain number of them *right* in succession In mental testing the corresponding concept would be to require the individual to get a certain number of items *wrong* in succession In general, any special technique designed for the analysis of learning data, such as a sequential type of analysis, would then be applicable in principle to the analysis of questionnaire and mental test data which also meet the abstract conditions of Method of Single Stimuli task A, monotone stimuli On the other hand, any technique designed for the analysis of such Quadrant IIa data is in principle applicable to the analysis of data from such learning experiments Hence, Lazarsfeld's latent distance model or latent structure analysis could be used to study learning

Lazarsfeld's system gains much of its importance from the fact that it is designed for the analysis of data collected by the Method of Single Stimuli, and this method is probably more widely used particularly in social science, than any other The method is widely used in public opinion surveys certain sociometric data also fall into this class of behavior When, for example, the members of a group are asked to indicate which individuals in the group they

prefer as friends, with no restriction on the number an individual picks, the behavior may be irrelative, task *A*, nonmonotone. If a restriction is placed on the number to be chosen, the behavior becomes relative (Method of Choice), task *A*, and the methods of analysis appropriate to Quadrant I are applicable.

BIBLIOGRAPHY

- 1 Bennett, J. F. *A method for determining the dimensionality of a set of rank orders* Ph D thesis, Univ of Michigan, 1951.
- 2 Bridgman, P. W. *The logic of modern physics* New York Macmillan, 1927.
- 3 Bronfenbrenner, U. A constant frame of reference for sociometric research, Part I *Sociometry*, 1943, 6, 363-397.
- 4 Carnap, R. *Foundations of logic and mathematics* Chicago Univ of Chicago Press, 1939.
- 5 Carroll, J. B. Problems in the factor analysis of tests of varying difficulty *Amer Psychologist*, 1950, 5, 369 Abstract.
- 6 Coombs, C. H. Psychological scaling without a unit of measurement *Psychol Rev*, 1950, 57, 145-158.
- 7 ———. Mathematical models in psychological scaling *J Amer Stat Assoc*, 1951, 46, 480-489.
- 8 ———. A theory of psychological scaling *Univ of Michigan Engineering Research Inst Bull No 34* Ann Arbor Univ of Michigan Press 1952.
- 9 Criswell, J. H. The measurement of group integration *Sociometry*, 1917, 10, 259-267.
- 10 Dushnik, B. Concerning a certain set of arrangements *Proc of the Amer Math Society*, 1950, 1, 788-796.
- 11 Festinger, L. The analysis of sociograms using matrix algebra *Hum Relat*, 1919, 2, 153-158.
- 12 Green, B. I. A general solution for structure analysis *Psychometrika*, 1956, 21, 1-16.

Class model of latent
1956

- 13 ——— Latent structure analysis and its relation to factor analysis
J Amer Stat Assoc 1952 47, 71-76
- 14 Guilford J P *Psychometric methods* New York McGraw Hill 1936
- 15 Gulliksen H *Theory of mental tests* New York Wiley 1950
- 16 Guttman L The basis for scalogram analysis In Stouffer S *et al*
Measurement and prediction studies in social psychology in World War II 4 Princeton Princeton Univ Press 1950 pp 60-90
- 17 Hevner K An empirical study of three psychophysical methods
J Gen Psychol 1930 4 191-212
- 18 Katz D Measurement of intensity In Cantril H *Gauging public opinion* Princeton Princeton Univ Press 1944 pp 51-65
- 19 Klingberg F Studies in measurement of the relations among sovereign states *Psychometrika* 1941 6 335-352
- 20 Lazarsfeld P The logical and mathematical foundations of latent structure analysis In Stouffer S *et al* *Measurement and prediction studies in social psychology in World War II 4* Princeton Princeton Univ Press 1950 pp 362-412
- 21 Likert R A technique for the measurement of attitudes *Arch of Psychol* 1932 No 140
- 22 Luce R D Connectivity and generalized cliques in sociometric group structure *Psychometrika* 1950 15 169-190
- 23 ——— and Perry A D A method of matrix analysis of group structure *Psychometrika* 1949 14 95-116
- 24 Moreno J L and Jennings H H Statistics of social configurations *Sociometry* 1937 38 1 342-374
- 25 Mosteller C F Remarks on the method of paired comparisons I The least squares solution assuming equal standard deviations and equal correlations *Psychometrika* 1951 16 3-9
- 26 Reese T W The application of the theory of physical measurement to the measurement of psychological magnitudes with three experimental examples *Psychol Monogr* 1913 55 No 3
- 27 Richardson M W Multidimensional psychophysics *Psychol Bull* 1938 35 659-660 Abstract
- 28 Saffir M A A comparative study of scales constructed by three psychophysical methods *Psychometrika* 1937 2 179-198
- 29 Stevens S S On the theory of scales of measurement *Science*, 1916 103, 677-680
- 30 ——— Mathematics measurement and psychophysics In Stevens

- S S (ed) *Handbook of experimental psychology* New York Wiley 1951, Chap 1
- 31 Stouffer, S, Guttman, L, Suchman, E A Lazarsfeld P F Star S A, and Clausen, J A *Measurement and prediction studies in social psychology in World War II, 4* Princeton Princeton Univ Press, 1950
- 32 Thomas, L G Mental tests as instruments of science *Psychol Monogr*, 1942, 54, No 3
- 33 Thurstone, L L A law of comparative judgment *Psychol Rev*, 1927, 34, 273 286
- 34 ——— Rank order as a psychophysical method *J Exper Psychol*, 1931, 14, 187 201
- 35 ——— *Multiple factor analysis* Chicago Univ of Chicago Press, 1947

Distribution-free Statistical Methods and the Concept of Power Efficiency

Keith Smith

Social scientists today are more aware perhaps than any other scientists of the restrictive nature of a priori assumptions concerning data. But although they sometimes follow tortuous routes to avoid "common sense" assumptions which go beyond the theoretical framework in which their studies are embedded, nevertheless, new measures and measuring devices are constructed and applied to two groups, and tests of the hypothetical difference between the groups are used which are based *entirely* on the assumption that the measured values are distributed according to the normal curve over both populations involved. Thus, although they are very sensitive to assumptions about what might be called their "real world," social scientists are prone to be insensitive to assumptions in the statistical systems in which they embed their data. The measurement or statistical systems into which data are mapped constitute an integral part of the theory and assumptions about the "real world."

For example, the crux of a sophisticated theoretical question about the "real world" may lie on a base which requires that, had the instrument been applied to all of some universe, usually hypothetical,

the distribution of measurements would have had a specific functional form. Ordinarily no evidence is ever gathered which might tend to confirm or deny this assumption.

This gives rise to two questions. Why is this so? What can be done about it? The first question is not too difficult to answer, at least in part. Statistics based on the normal distribution have had tremendous success in both the physical and the biological sciences. Furthermore, if the normality assumption is justified, no bothersome decisions need to be made concerning which test of a specific hypothesis to use. The statistician has been able to tell us which test is "best" (in a sense to be discussed later). A third reason—a reason for which the statistician need not take all the blame—is, according to the statistician Geary (4), "the beauty of the mathematical theory and the facility of algebraic manipulation involved." The social scientist has been all too prone to seek the approval (and gratitude) of the statistician by allowing the normality assumption to be made, even when knowledge of the subject matter involved indicates that the assumption is invalid.

In recent years mathematical statisticians have begun to construct answers to the second question, "What can be done about it?" This area of statistics must be one not inextricably bound to normal theory. One such area is called variously "order statistics," "non-parametric statistics," or "distribution-free" statistics. Even here not all assumptions concerning the mathematical form of the distributions under consideration have been dropped. One must still assume for most methods that the population is continuous. Although this is still a very strong assumption, its palatability is increased by the absence of the additional assumptions required by conventional or parametric statistics.

Statistical procedures, both parametric and nonparametric, fall into two classes corresponding to two general purposes for which they may be used: (1) testing whether a population (S) from which one or more samples were drawn has a certain characteristic (e.g., Is the population normal? Do the populations have the same means?), and (2) estimating some number characteristic (parameter) of the population represented by a sample (e.g., Within what limits does the mean probably lie?). These are ordinarily called *test statistics* and *estimation statistics*, respectively.

A number of different statistical procedures have been devised to accomplish each of these purposes. Some of the procedures are alternatives under a given set of conditions. It becomes desirable, then, to be

able to evaluate or compare procedures in order to be able to select one that in some sense best satisfies the experimenter's objectives

The comparison of *test statistics* is based on the question "How often will it give the right answer?" For example, if a difference exists between the means of two populations, will one test more often yield a significant difference than another test for random samples of the same size? The comparison of *estimation statistics*, succinctly referred to as *estimators*, is based on two questions. How often will it give the right answer, and how much information does it give about the population parameter?

The "How often" question is answered in the same way for both classes of statistical procedures by means of what is called the *power curve*. This will be discussed in general terms in the second section of the chapter. In the third section, where various nonparametric tests of significance are discussed, the concept of the power curve will be used for the purpose of comparing the nonparametric tests with conventional parametric tests.

The second question used to evaluate estimators, "How much information does it give about the population parameter?" pertains to the size of the confidence interval which brackets the parameter. Other things being equal, the shorter the confidence interval an estimator yields on the average, the better the estimator. The fourth section of the chapter will deal with estimators, unfortunately, however, very little has yet been accomplished on the problem of comparing nonparametric estimators.

CRITERIA FOR THE COMPARISON OF STATISTICAL TESTS

As a consequence of the fact that alternative statistical procedures have been developed which *can* be applied to the same data, it becomes both necessary and desirable to provide a rational basis for choosing one procedure instead of another. In this section, a number of criteria for making this decision will be presented, with particular emphasis on the concepts of the power and the power efficiency of a statistical test. This choice of emphasis is due to the general unfamiliarity of these concepts to social scientists and to the significant role they should play in the choice of a statistical test.

Before the criteria are presented and discussed, it will be desirable

to review briefly certain characteristics of statistical tests in general. In this presentation, a knowledge of elementary sampling theory will be assumed.

Definition of Statistical Tests

A statistical test is a formal mechanism, based on probability, for arriving at a decision about the reasonableness of an assertion. The assertion is called a hypothesis, and any value (usually a number) obtained from a sample of data is a test statistic. The mechanism makes use of the one or more obtained values (test statistics) to arrive at a probability statement about the assertion. But the mechanism almost always makes use of more than that, and these additional things are other assertions about the population from which the sample was drawn (e.g., normally distributed), the manner in which the sample was drawn (e.g., randomly), etc.

From the point of view of the person making use of the statistical test, the assertions involved are of two kinds. One is an assertion directly related to the purpose of the investigation: this is an assertion which is to be tested and is called a hypothesis. All the other assertions are those which it is necessary to assume to make a probability statement. This second set of assertions is called the *model*. All probability statements about a hypothesis are preceded, implicitly or explicitly, by the qualifier "If the model used was correct, then . . ."

It should be clear that the weaker the assertions that define the model, either by virtue of fewer assumptions or less restrictive assumptions, the more general the conclusions. On the other hand, the stronger the model—the more assumptions built into it—the more powerful will be the test of the hypothesis.

As an illustration, a statistical hypothesis might be "Population *A* has a mean equal to the mean of population *B*." Such a hypothesis contains no normality statements and, in fact, no statements about any characteristics of the populations other than the means. If a test of this hypothesis is used which contains an assumption of normality in the parent populations, normality becomes a part of the model and normality is *not* tested.

Two characteristics of a statistical test have been discussed: (1) the model, the assertions which are assumed to be right, and (2) the statistical hypothesis, the assertion which is to be tested, called a null hypothesis. A third characteristic always required in a statistical test is

the hypothesis or class of hypotheses which are alternative to the null hypothesis. These alternative hypotheses are *always* part of the test, whether explicit or implicit. These constitute the assertion that is accepted if the null hypothesis is rejected.

From this point of view, it is evident that there are two types of error which are possible in arriving at a decision about the null hypothesis. *Type I* accepting the alternative hypothesis when the null hypothesis is true. *Type II* accepting the null hypothesis when it is false.

There is an inverse relationship between the likelihoods of making these two types of error. If the null hypothesis were always accepted regardless of the sample test statistic, the probability of a Type I error would be zero, but there would be a maximum likelihood of a Type II error. Correspondingly, if the null hypothesis were always rejected, the danger of a Type II error would vanish, but there would be a maximum likelihood of a Type I error. A compromise must be reached between these two dangers, and various statistical tests offer the possibility of different compromises. This is the problem for which the concepts of the *power* of a statistical test and its *power efficiency* are relevant.

Power and Power Efficiency

The concepts of power and power efficiency will be introduced by way of an illustration.

A COMPARISON OF TWO STATISTICAL TESTS A common problem is a test of a hypothesis about the value of the mean of a population. The null hypothesis might be that the mean of the population from which the sample was drawn is 4. An alternative hypothesis might be that it is 5, and an infinite class of alternative hypotheses all used at once might be that it is greater than 4.

The simplest possible case of a statistical test is whether a population has one or the other of two specific values, all the other characteristics of the population having been specified in the model. Let the assertions that the population is normal with variance equal to one constitute the model. Let the null hypothesis be

$$H_0 \text{ the mean is } 1$$

Let the alternative hypothesis be

$$H_1 \text{ the mean is } 2$$

A sample of size 1 is drawn and, on the basis of the model and the test statistic (the sample value), a choice is to be made between the null hypothesis and its alternative. The situation specified by the model is illustrated in Figure 1.

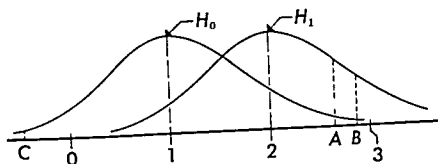


FIG 1 The distributions specified by the alternatives H_0 and H_1

Let us hold constant the probability of a Type I error at 5 percent. Two statistical tests that may be used are the one-tailed test and the two-tailed test. The probability of a Type I error is called the significance level of the test.

To use the one-tailed test, we shall choose a number A larger than 1 (the mean under the null hypothesis). If our sample value is larger than A , we shall agree to reject the null hypothesis and, of course, to accept the only alternative hypothesis—i.e., that the mean is 2. If the sample value is less than A , we shall agree to accept the null hypothesis. A must be chosen in such a way as to make the probability of a Type I error 0.05. This will be the case if the area under the curve H_0 to the right of A is 0.05, and from a table of normal curve areas we see that A must be 2.645. The probability of making a Type II error is the area under the curve H_1 and to the left of A —i.e., the proportion of times we will accept H_0 when H_1 is really true. Again, from a table of normal areas, we see that this probability is 0.74.

To use the two-tailed test, we choose two numbers, B and C , equally distant from the mean under the null hypothesis. If our sample value is farther from 1 than these numbers, we agree to reject the null hypothesis and accept the only alternative. If the mean is nearer 1 than B and C are, we accept the null hypothesis. The area under H_0 to the right of B must be 0.025, and the area to the left of C must also be 0.025 to make the total 0.05. And from our table $B = 1 + 1.96 = 2.96$ and $C = 1 - 1.96 = -0.96$. The probability of a Type II error is,

of course, the area between -0.96 and 2.96 under the curve H_1 . This area is 0.83 .

Thus, although the two tests have the same amount of Type I error, the two-tailed test of this hypothesis has more probability of a Type II error, and in this sense is the poorer of the two tests against this alternative. In this simple example, it should be apparent that values to the left of C are poor evidence for H_1 . The two-tailed test is used here only to illustrate the fact that the rejection level of a test is not a sufficient basis for its evaluation.

POWER We are now ready to introduce the concept of power. The power of a statistical test against a specific set of alternative hypotheses at a specific significance level is given by the equation

$$\text{Power} = 1 - (\text{Probability of a Type II error})$$

Power might alternately (but equivalently) be defined as the probability of rejecting the null hypothesis when the alternative hypothesis is true. In Figure 1, the power of the one-tailed test is the area to the right of A under the curve H_1 . In tests of means from normal populations the concept of power is synonymous with the Fisher concept of "amount of information" (3).

Research workers sometimes use certain tests because the tests are "conservative," meaning that they are less powerful than others that might be used. The discussion above exposes this as a rather peculiar sort of conservatism. Testable theoretical deductions are rare enough without loading the dice against them, but this is what the research worker does if he wishes to reject the null hypothesis. Furthermore, the "conservative" test is in no sense conservative if the research worker wishes *not* to reject the null hypothesis (as in homogeneity of variance tests in analysis of variance).

In our example above, the power of the one-tailed test is 0.26 , whereas the power of the two-tailed test is 0.17 . Thus, the one-tailed test is "more powerful" than the two-tailed test at the 5-percent level against the hypothesis that the mean of a normal distribution with variance 1 is 2, when the mean under the null hypothesis is 1. So far, all these qualifying phrases are necessary and illustrate the caution essential when statistical tests are compared. It can be shown, in this simplest of cases, that, at any significance level, whenever all alternative hypotheses state means on one side of the null hypothesis mean, the one-tailed test will be more powerful than the two-tailed test.

To show that the power of a test is determined by the alternative hypotheses, let us take as another example the same situation except that here the alternative hypothesis is that the mean is 3 (Fig 2). In this case the values of A , B , and C remain the same, being determined by the null hypothesis. The area to the left of A under H'_1 here is 0.36, whereas the area under H'_1 between B and C is 0.52. Consequently, the power of the one-tailed test is $1 - 0.36 = 0.64$, whereas the power of the two-tailed test is $1 - 0.52 = 0.48$.

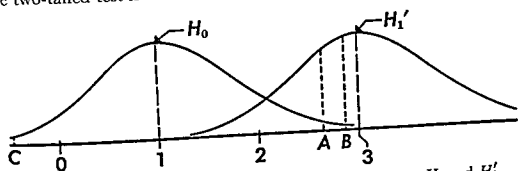


FIG 2 The distributions specified by the alternatives H_0 and H'_1

One can plot, for this model and significance level, the power of each test against a hypothesis versus the value stated in the hypothesis. The curves obtained will be the "power curves" for the one-tailed and two-tailed tests (Fig 3).

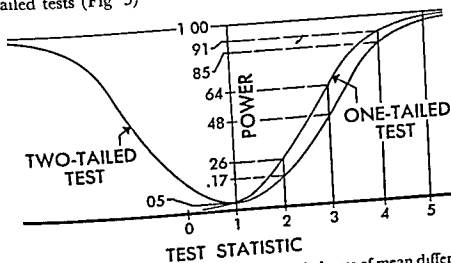


FIG 3 Power curves of the one- and two-tailed tests of mean difference

It is, of course, not necessary that there be but one alternative. Another test using the same model could have the two alternatives

H_1 mean is 2, H_2 mean is 3 Here, also, the one tailed test is more powerful against both alternatives, since it is more powerful against each singly The difference in tests would only be that rejecting H_0 would not imply accepting a single alternative but would imply only accepting the pair of alternatives

However, if the alternatives had been H'_1 mean is 0, and H_2 mean is 2, a considerable difference obtains As Figure 3 indicates, the one-tailed test has almost no power (0.004) against the hypothesis that the mean is zero, whereas the two tailed test has a power of 0.17, at least, against both alternatives

A still more complex test would be one with the infinite number of alternatives H_1 mean is greater than 1 Again, the two-tailed test is less powerful than the one-tailed test, since it is less powerful for all alternatives

Last, but most often used, is the test with the infinite class of alternative hypotheses the mean is *not* 1 As would be expected from Figure 3, the two tailed test would be used, since it is only slightly less powerful on the right of the null hypothesis, but extremely more powerful on the left

POWER AND SAMPLE SIZE The preceding illustration was entirely in terms of a sample of size one The illustration can be generalized by using, instead of the sample value, the number $(\bar{X}\sqrt{n})/\sigma$ where σ is specified in the *model* as the standard error of the normal distribution involved, n is the sample size, and \bar{X} is the sample mean As n increases the test statistic, $(\bar{X}\sqrt{n})/\sigma$ increases and, by Figure 3, the power of

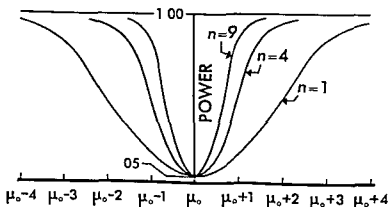


FIG. 4 Power curves of the two-tailed test with varying sample size

the test increases also. Figure 4 illustrates this increase in power of the two tailed test of the mean for samples of size 1, 4, and 9 from normal populations with unit variance, and μ_0 is the mean under the null hypothesis.

For almost every statistical test in use, including all those to be discussed in this chapter, increasing sample size increases the power of a statistical test.

To summarize, (1) the power of a statistical test is the probability of rejecting the null hypothesis when it is false, (2) power is relative to the model employed and to the alternative hypotheses (possibilities) entertained, (3) as a general rule, the power of a statistical test increases with sample size.

POWER EFFICIENCY Unfortunately, there are considerations other than those of power that must be made in the choice of a statistical test. Is the test simple computationally? Is the model (set of assumptions) required for this test "true to life"? Leaving the first question for the moment, let us consider how the second might be answered.

It was stated earlier that the weaker the assertions constituting the model, the more general the conclusions but the less powerful the statistical test. This is very generally true for a given sample size but not necessarily so if the tests use different sized samples. Test A may be better (more powerful) than test B for samples of size twenty, but test B may very well be more powerful with a sample of size twenty than test A is with a sample of size ten. In other words it is necessary to pay for increased generality of conclusions with a larger sample. Power efficiency is a measure of how much one has to pay in any specific case.

If test A with a sample size N_A has the same power as test B with the sample size N_B , where test B is the most powerful test (known or hypothetical) for N_B observations, then test A has power efficiency $100(N_B/N_A)$ percent. If test A requires a sample of 10 to be as powerful as test B with a sample of size 6, then test A has power efficiency of 60 percent. The question becomes "Is it worth the extra expense of taking a sample this much larger to arrive at more general conclusions?"

The preceding paragraph is adequate when there is only one alternative hypothesis. When there are more than one, the phrase "as powerful as" loses meaning. Test A may be more powerful against alternative one but less powerful against alternative two. How, in this instance, can we say one test is more powerful than the other, or as powerful as the other? In Figure 5, which test is more powerful?

No unique or "best" answer exists. Many rational answers can be found, two of which can be discussed here. A commonly used answer is the equal-area criterion. In Figure 5, test *A*, by this criterion, would be as powerful as test *B* if the ruled area were equal to the cross-hatched area. Another means of equating power curves is to adjust sample sizes so that the power curves intersect at some central power level (say 0.50). The two definitions do not yield essentially different answers ordinarily, and the two tests in Figure 5 are a case in point.

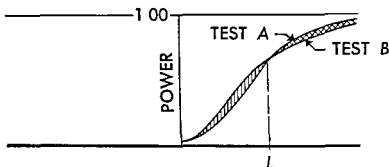


FIG. 5. A portion of the power curves of two hypothetical tests.

A second and last important complication is that power efficiency itself may vary for different sized samples. Test *A* may have a power efficiency of 0.70 with samples of 5 and a power efficiency of 0.95 with samples of 100. A test's power may increase rapidly with sample size while the power of the most powerful test is increasing only slowly.

To summarize, then, power efficiency is a *measure* of the power of a statistical test relative to the most powerful test possible. Although power efficiency may be defined in several ways, the definitions are roughly equivalent. Finally, power efficiency is not completely independent of sample size in all cases.

TESTS OF HYPOTHESES WITH ORDER STATISTICS

The rest of this chapter will be devoted to a discussion of a number of nonparametric tests and estimation procedures. All available information concerning the power or power efficiency of each test will

be presented. The tests, in almost all cases, will be compared with tests based on normal models. Unfortunately, the power and the power efficiency of a test cannot be determined unless some definite functional form for the underlying distributions is specified.

For several reasons the tests will be compared entirely on the basis of underlying normal distribution—that is, the power efficiencies of nonparametric tests relative to the corresponding normal tests will be given for those cases where the normal tests are strictly applicable. We shall be answering the question “How much do we lose by using these nonparametric statistics when we *could* use normal statistics?” “How much could our apprehensions cost us if our fear (of non-normality) was unjustified?”

The more pressing question, “How important is it that my results apply generally rather than to populations distributed normally?” must, of course, remain unanswered, since it can only be answered by the research worker each time he considers a test of experimental data.

The experimental tests presented will fall into two natural but not mutually exclusive groups which for convenience will be labeled *tests of location* and *tests of relation*. These correspond, respectively, to the ordinary tests of mean difference and tests of correlation, with tests analogous to the analysis of variance occupying some middle category.

Tests of Location

THE SIGN TEST This test is useful in cases where the t test of differences is ordinarily used, that is, where a set of paired observations is available. A common application is the before after type of experiment, in which measurements are made on each subject some treatment or stimulus is applied to each, and the measurements are repeated. The null hypothesis H_0 would be that there is no effect, whereas the alternatives could be either H_1 the effect is positive (negative), or H'_1 there is an effect. For the purposes of the sign test, these are reformulated in this way:

H_0 The median of the distribution of differences between before and after measures is zero. against H_1 this median is positive (negative), or H'_1 this median is not zero.

The power efficiency (in the remainder of this chapter power efficiencies are quoted for 5 percent level tests) of this test in normal

samples is rather low, ranging from 63.7 percent for larger samples up to 68 or 70 percent for small samples. The sign test with 18 pairs of observations is approximately equivalent to a t test with 12 pairs when the t test is applicable, 40 pairs for the sign test compared with 28 or 29 for the t test.

Let us suppose that a questionnaire about attitudes toward Negroes has been administered to a group of 22 subjects, after which the subjects undertook an extensive study of race prejudice. The following results were obtained on a follow-up questionnaire.

<i>Subject</i>	<i>After</i>	<i>Before</i>	<i>Difference</i>	<i>Subject</i>	<i>After</i>	<i>Before</i>	<i>Difference</i>
1	35	31	+ 4	12	28	27	+ 1
2	36	29	+ 7	13	27	26	+ 1
3	29	35	- 6	14	27	25	+ 2
4	34	32	+ 2	15	33	32	+ 1
5	33	29	+ 4	16	42	40	+ 2
6	28	33	- 5	17	19	18	+ 1
7	33	30	+ 3	18	37	36	+ 1
8	28	38	- 10	19	40	40	0
9	28	35	- 7	20	32	31	+ 1
10	25	22	+ 3	21	31	27	+ 4
11	33	31	+ 2	22	30	29	+ 1

If the null hypothesis were true—that is, if these 22 differences were drawn from populations with a median of zero—we should expect about half of them to be positive and half negative. We see that, in fact, 17 are positive, 4 negative, and one is zero. We should expect that the distribution of +’s and -’s would be about the same as that of heads and tails in tossing 22 unbiased coins. The probability of 17 heads and 4 tails (*one coin rolling out of sight*) is less than 0.01, so we would reject H_0 at the 1-percent level if our alternative hypotheses are specified by H'_1 . Had we decided *beforehand* to use the sign test with the one-tailed alternatives H_1 , the effect is positive, we would have rejected H_0 at the $\frac{1}{2}$ -percent level.

The t test of differences of either the one-tailed or two-tailed form here would have yielded a quite insignificant result ($t_{21df} = 0.60$) which would have ignored the fact that although the differences were small, they were almost all in the same direction.

Probability levels for the sign test are tabulated in Dixon (1). The

5-percent level of the two-tailed test can be obtained from the formula $[(N - 1) / 2] - (0.98) \sqrt{N + 1}$, the level being the integral part of this value, where N is the number of pairs in the sample, ignoring all pairs with a difference of zero. In our sample, $N = 21$, the formula yields $(20/2) - (0.98) \sqrt{22} = 10 - 4.596 = 5.404$ and we shall reject the null hypothesis at the 5-percent level if there are 5 or fewer plus or minus signs among the 21 differences.

The sign test is perhaps the simplest of all distribution-free tests and at the same time is most readily generalized to analysis of variance-type problems. This generalization will be discussed after the presentation of a number of two-sample tests.

THE WALD-WOLFOWITZ RUN TEST (16) The Wald-Wolfowitz Run test is specifically designed to test the null hypothesis H_0 : two samples were drawn from populations having the same continuous distribution. The alternatives are the extremely large class characterized by H_1 : two samples were drawn from *different* continuous distributions. This test, then, would tend to reject the null hypothesis if two populations were different in any respect.

The assumption of continuity is a critical assumption here. It implies that gross measurement is the only reason for two observations to have the same value.

The power efficiency of this test is not known. However, a small number of empirical tests by the author have indicated a number somewhere in the neighborhood of 75 percent in normal samples when the distributions differed only in means and when sample sizes were about 20.

Let us say that we have two groups of high school students in a large city, one of which has just completed a course in religious education, the other group not having had this course. We wish to test the hypothesis that our religious-education course had no effect on attitude toward religion as measured by a new scale. Say, also, that the scores on the attitude scale are as follows:

Religious education 52, 53, 71, 86, 95, 108, 115, 120, 141, 152,
165, 218

No religious education 30, 45, 54, 74, 75, 81, 101, 104, 146, 151,
170, 171

The test consists in arranging all the scores of both groups in order and then counting the number of runs of scores coming from each group. In the example we have 30, 45, 52, 53, 54, 71, 74, 75, 81, 86,

95, 101, 104, 108, 115, 120, 141, 146, 151, 152, 165, 170, 171, 218 Our test statistic is the number of runs, r , which in this case is 12 Under the null hypothesis, r_0 is given by

$$r_0 = E(r) = (2 N_1 N_2) / (N_1 + N_2) + 1$$

The variance of r , $\sigma_r^2 = \frac{2 N_1 N_2 (2 N_1 N_2 - N_1 - N_2)}{(N_1 + N_2)^2 (N_1 + N_2 - 1)}$, where N_1 is the number in one sample, N_2 the number in the other
In the example $N_1 = N_2 = 12$,

$$r_0 = \frac{2 (12) (12)}{24} + 1 = 13$$

$$\sigma_r^2 = \frac{2 (12) (12) (264)}{(24)^2 (23)} = 5.74$$

$$\sigma_r = 2.40$$

For values of N_1 and N_2 larger than 10, r is approximately normally distributed and large values of r tend to confirm the null hypothesis Hence we form the normal deviate $C = (r - r_0) / \sigma_r$ and reject the null hypothesis at the 95 percent level whenever $(r - r_0) / \sigma_r < -1.645$

In this example, $C = -1/2.40 = -0.417$ and the null hypothesis would not be rejected Our conclusion would be that we have no evidence that such a course has any effect on attitudes toward religion

Had we assumed normality and used the two tailed t test for significance of difference between means, we should have reached about the same conclusion ($t = 0.73$), except that it would have been qualified by the clause "if scores on this scale are normally distributed"

It is important to note again, that, although we used a normal deviate to find the probability of our observation, the validity of the test depends in no way on the normality of the observations but only on the distribution of r in two samples from the same population That this distribution becomes normal for large samples can be proved mathematically We use it here as an approximation to the true distribution, which is much more difficult to use For sample sizes of 12 and 12, the approximation would call 9 runs significant at the 5-percent level, whereas the true distribution would call 9 runs significant at the 6.9 percent level The correspondence would be even closer for larger samples¹

¹For exact significance levels for $N_1, N_2 < 10$, see Swed and Eisenhart (14)

Theoretically, if the original population is continuous, there should be no two observations with the same value. Because of coarse measurement, however, such ties do occur. When they occur, the sequence of ranked observations is not unique. Any of several sequences may adequately describe the data, and there is no logical way of distinguishing between them. Of course, if all ties are within one sample, the number of runs is the same for each sequence and our test is unaffected, but if observations in one sample are tied with observations in the other, the number of runs cannot be uniquely determined from the data.

Let us consider our original example, slightly altered² so that there is one tie across samples. The combined, ordered sample might be 30, 45, 52, 52, 53, 54, 71, 74, 75, 81, 86, 95, 101, 104, 108, 115, 120, 141, 146, 152, 165, 170, 171, 218. Here $r = 12$ or 14, depending upon which observation of 52 is in reality the smaller. Neither value would be significant, and consequently our decision would be the same. If the proportion of ties is very large, ordinarily the number of runs is very indeterminate, often ranging from very significant to very insignificant numbers, and in such cases this test is inapplicable.

THE MANN-WHITNEY OR WILCOXEN TEST (6, 18) Although subject to the same ambiguities when ties in rank occur in the data, the Mann-Whitney or Wilcoxen test is valuable since it provides a one-tailed set of alternatives. The null hypothesis H_0 is again that the two samples are drawn randomly from populations F and G having the same distribution. Since this, like all nonparametric tests, is a test of distributions, the alternatives must be stated rather differently. The statement of the alternatives is H_1 : F is stochastically larger than G . If λ is one observation from the distribution F and Y is from the distribution G , F is stochastically larger than G if the probability of λ larger than Y is greater than $1/2$. Loosely speaking, this implies that the "bulk" of F is farther to the right than the "bulk" of G , or that F accumulates more "slowly" from the left than does G . If the two populations are normal with the same variance, this implies that the mean of F is larger than the mean of G .

The Mann-Whitney test makes exactly the assumptions of the Wald-Wolfowitz test.

Van der Vaart (15) has investigated the power function of the Mann-Whitney test for normal samples and has found that, although the t test is more powerful, the differences in power are quite small for

² An observation of 52 was added to the "No religious education" group

small samples and not very great even for larger samples. This test is probably the simplest and most powerful nonparametric test yet devised for detecting differences in location. The Wald-Wolfowitz test seems to be better for detecting differences in dispersion.

As an illustration of the use of this test, suppose that from our previous example our alternative hypotheses had been H_1 : students who have taken religious education have higher scores on this attitude scale than do other students. Our statistic U is given by the number of students without religious education who have higher scores than a specific student with the education. This number is then summed over all students with religious education. Let us tabulate this statistic for our example.

TABLE I

<i>Score on attitude scale for religious-education students</i>	<i>Number of non religious-education students higher</i>
218	0
165	2
152	2
141	4
120	4
115	4
108	4
95	6
86	6
71	9
53	10
52	10

$$U = 61$$

If the null hypothesis is true, the expected value of U_0 is $\frac{1}{2} N_1 N_2$, or 72.

The variance is $\sigma_U^2 = 1/12 N_1 N_2 (N_1 + N_2 + 1)$ or 300. When the two samples are both larger than 10, $(U - U_0)/\sigma_U$ is approximately normally distributed, and we shall reject the null hypothesis at the 5-percent level if this is less than -1.645 . In our case, $C = -11/(10\sqrt{3}) = -0.635$, which would have been significant at the 26 percent level.

Ties in rank have less effect on U than on r , but large numbers of ties still vitiate the value of the test. When ties do occur, they should be counted as $\frac{1}{2}$ point.

MARSHALL'S TEST (7). Marshall's Test is a large-sample test for exactly the same purpose as the preceding test. It has the additional important feature that ties in rank are quite unimportant in its application. Also important is the fact that it may be applied to grouped data for two samples if each sample is measured on the same continuum. It is to be expected that this test is very useful to the survey researcher, whereas the preceding test is more useful to the experimenter.

The power efficiency of the Marshall test for extremely large samples has been investigated for normal samples with known variances. Preliminary results indicate that the power efficiency of the test is larger than 0.90 if the number of class intervals is larger than 5.

Marshall's original paper (7) contains an interesting application which should be read as well as the one presented here. The data presented here are taken from records of the 1937 American Council Psychological Examination. The sample size is extremely large, but use of the test requires only that each interval contain at least a total of 10 observations in the two samples. The hypothesis to be tested will be H_0 : males and females have the same distribution of scores on the 1937 ACPE, against the class of alternative hypotheses H_1 : males have higher scores than females. The original data are contained in Table II.

TABLE II

<i>Scores</i>	<i>Male</i>	<i>Female</i>	<i>Total</i>
		27	59
0-29	32	387	970
30-59	583	1,606	3,667
60-89	2,061	3,439	7,603
90-119	4,164	5,090	11,035
120-149	5,945	5,376	11,912
150-179	6,536	4,440	9,892
180-209	5,452	3,153	6,972
210-239	3,819	1,832	4,059
240-269	2,227	789	1,905
270-299	1,116	248	681
300-329	433	60	183
330-359	123	3	12
360-...	9		58,950
	32,500	26,450	

Each column of the original table is then cumulated in Table III to provide the data from which we work.

TABLE III

Group	Scores	Cumulative male	Cumulative female	Cumulative total	\hat{p}_s	\hat{q}_s	$\sum_{k=s+1}^{13} \hat{q}_k$	$\sum_{k=s+1}^{13} \hat{p}_s \hat{q}_k$	$\hat{p}_s \hat{q}_s$
1 =				59					
1	0-29	32	27		0.0010	0.9990	4.1180	0.00412	0.00100
2	30-59	615	414	1,029	0.0174	0.9826	3.1354	0.05456	0.01710
3	60-89	2,676	2,020	4,696	0.0796	0.9204	2.2150	0.17631	0.07326
4	90-119	6,840	5,459	12,299	0.2086	0.7914	1.4236	0.29696	0.16509
5	120-149	12,785	10,549	23,334	0.3957	0.6043	0.8193	0.32420	0.23912
6	150-179	19,321	15,925	35,246	0.5978	0.4022	0.4171	0.24934	0.24044
7	180-209	24,773	20,365	45,138	0.7655	0.2345	0.1826	0.13978	0.17951
8	210-239	28,592	23,518	52,110	0.8838	0.1162	0.0664	0.05868	0.10270
9	240-269	30,819	25,350	56,169	0.9526	0.0474	0.0190	0.01810	0.04515
10	270-299	31,935	26,139	58,074	0.9849	0.0151	0.0039	0.00384	0.01487
11	300-329	32,368	26,387	58,755	0.9965	0.0035	0.0004	0.00040	0.00349
12	330-359	32,491	26,447	58,938	0.9996	0.0004	0.0000	0.00000	0.00040
13	360-	32,500	26,450	58,950	1.0000	0.0000	0.0000	0.00000	0.00000
Totals		255,747	209,050					1.32630	1.08210

The final five columns are all computed on the basis of the Cumulated Total column. The column \hat{p}_i is the cumulated percentage of the total (for example, in the third row $0.0796 = 4696/58,950$),

\hat{q}_i is one minus \hat{p}_i , each $\sum_{k=i+1}^{13} \hat{q}_k$ is the sum of the two entries immediately below it in columns $\sum_{k=i+1}^{13} \hat{q}_k$ and \hat{q}_i , the headings on the last two

columns are self-explanatory. Where λ is the sum of the cumulated frequencies in the hypothesized larger sample, m the number in the sample, Y the sum of the cumulated frequencies in the other sample, and n the number in its sample, the test statistic is

$$S = (Y/n) - (X/m)$$

The variance of S is

$$\sigma_s^2 = [(1/m) + (1/n)] \left[\sum_{i=1}^j \hat{p}_i \hat{q}_i + 2 \sum_{i=1}^{j-2} \sum_{k=i+1}^{j-1} \hat{p}_i \hat{q}_k \right],$$

where j is the number of categories

Our test statistic here is $S = (209,050/26,450) - (255,747/32,500) = 7.9036 - 7.8691 = 0.0345$. The variance of S , $\sigma_s^2 = [(1/32,500) + (1/26,450)] [1.0821 + 2(1.3263)]$, where 32,500 and 26,450 are the sample sizes, and 1.0821 and 1.3263 are the sums of the last and next-to-last columns

$$\sigma_s^2 = (0.68576 \times 10^{-4}) (3.7347) = 0.00025611, \text{ and } \sigma_s = 0.0160$$

S is approximately normally distributed, and the normal deviate in this case is $C = 0.0345/0.0160 = 2.156$. The significance level obtained from the upper tail of the normal curve for our result is 0.016, between the 1- and 2-percent level.

SMIRNOV TEST (8) The Smirnov test is based upon the same reasoning as the Marshall test. It is a good deal simpler computationally, tests exactly the same hypotheses, and also is valid for large samples only. Its use is not recommended over the Marshall test unless time is an important factor.

Little is known of the power efficiency of the Smirnov test, except that it is almost certainly less power efficient than the Marshall test. It is believed to be more powerful than the chi-square goodness-of-fit tests (8).

We shall use the same data contained in Table III to illustrate the application of this test. The numbers needed will be the sample sizes, 32,500 and 26,450, and the cumulative percentages for each sample as contained in Table IV.

TABLE IV

<i>Scores</i>	<i>Cumulated percentage, female</i>	<i>Cumulated percentage, male</i>	<i>Difference</i>
0-29	0 00102	0 00098	0 00004
30-59	0 01565	0 01892	-0 00327
60-89	0 07637	0 08234	-0 00597
90-119	0 20639	0 21045	-0 00406
120-149	0 39883	0 39338	0 00545
150-179	0 60207	0 59449	0 00758
180-209	0 76994	0 76224	0 00770
210-239	0 88915	0 87975	0 00940
240-269	0 95841	0 94827	0 01014
270-299	0 98824	0 98261	0 00563
300-329	0 99761	0 99593	0 00168
330-359	0 99988	0 99972	0 00016
360-	1 00000	1 00000	0 00000

Our test statistic is the maximum difference of cumulative percentages, which is 0 01014 in this case, multiplied by a factor which takes into account the sample size, $\sqrt{nm/(n+m)}$, which in our case is 120 75. The product of these, λ , is 1 2244. The probability level is $e^{-2(1.2244)^2}$ or approximately 0 05. In general, the 5-percent level of λ is 1 224, the 1-percent level is 1 517. The significance level must be interpreted somewhat differently here, since it is in reality an upper bound to the true significance level. Had we computed λ using the largest difference from all possible classifications, we should have found the true significance level which would, of course, have been at least as small as the one we obtained here.

It is rather interesting to note that the normal test of difference between means, which is probably applicable to these extremely large samples, yields a difference in means of 1 033 with a standard error of 0 474, which is significant at about the 1 5-percent level for a one tailed test. The correspondence with the Marshall test is amazingly close.

PITMAN'S RANDOMIZATION TESTS (11) Pitman's Randomization tests are a set of nonparametric tests which are applicable in a wide

number of situations but which, unfortunately, are much too difficult computationally to be used often. The tests are less difficult for extremely small samples ($n < 10$).

Let us suppose, as an example, that we have the responses of two groups of five people each to the question "How much time per week should a man spend doing home repairs?" We wish to test the hypothesis of no mean difference against the alternative set of hypotheses that the Group 1 population has a different mean from the Group 2 population. Let the data be as follows

<i>Person</i>	<i>Group 1</i>	<i>Group 2</i>	<i>Person</i>
(1)	15 20	14 45	(6)
(2)	20 10	6 30	(7)
(3)	7 95	10 90	(8)
(4)	10 80	8 10	(9)
(5)	6 85	8 40	(10)
$\bar{X}_1 = 12.18 \quad \bar{X}_2 = 9.63$		$\bar{X}_1 - \bar{X}_2 = 2.55$	

Now if the null hypothesis is true, all ten observations are from a common population, and the splitting of the sample in two is merely a matter of chance—i.e., any of the $\binom{10}{5} = 252$ ways of carrying out such a split is equally likely. On the other hand, if the alternative hypothesis is true, we expect that there would be more than a chance difference between the group means. Only once in 252 such experiments should we expect the five largest observations in one group and the five smallest in the other group. If such a thing occurred, we should reject the null hypothesis at the 0.4-percent level. Similarly, the observed mean difference would be among the twelve largest possible such differences less than 5-percent of the times if the observations are from a single population—i.e., if the grouping is random. Here, then, is our test. We simply attempt to find 12 groupings of the 10 observations which give a larger mean difference than the one (2.55) we found. If the one we found is among the twelve largest, we reject the null hypothesis at the 5-percent level. If it is not, we do not reject it.

In our example, we cannot reject the null hypothesis. More than

twelve combinations which give a larger mean difference are shown below

<i>Persons in Group 1</i>	\bar{X}_1	$\bar{X}_1 - \bar{X}_2$
1, 2, 6, 8, 4	14.2	6.59
1, 2, 6, 8, 10	13.72	5.63
1, 2, 6, 8, 9	13.66	5.51
1, 2, 6, 8, 3	13.63	5.45
1, 2, 6, 8, 5	13.41	5.01
1, 2, 6, 8, 7	13.30	4.79
1, 2, 6, 4, 10	13.70	5.59
1, 2, 6, 4, 9	13.69	5.57
1, 2, 6, 4, 3	13.61	5.41
1, 2, 6, 4, 5	13.39	4.97
1, 2, 6, 4, 7	13.28	4.75
1, 2, 6, 9, 10	13.25	4.69
1, 2, 6, 10, 3	13.22	4.63

Under rather general conditions, this test leads to the same conclusions as would the ordinary t test, but the power efficiency of the Pitman tests in nonnormal universes is unknown.

THE MEDIAN TEST (17) The median test is rather similar to the run test presented earlier and is a generalization of the sign test. It is subject to somewhat less difficulty with ties in rank than is the run test, since ties are important only when they occur at the median of the combined sample. It has the further advantage of being rather insensitive to differences in dispersion. If, for example, the ordered samples A and B took the form $aaaabbbbbbbbaaaa$, where a is an observation from the A sample and b an observation from the B sample, the run test would reject the null hypothesis of no difference in the two populations, since there are only three runs. Since the samples have the same median, the median test would not be significant. The difference in the two samples is the wide dispersion of A compared to the dispersion of B . The hypotheses tested here are H_0 the two populations have the same median, against H_1 the medians are different.

Results on the power efficiency of this test are not available. As was noted above, it should be more powerful against alternatives which

specify differences in location than against alternatives specifying differences in dispersion

In application, the test is particularly simple. Let us say that we have n_1 observations in one sample and n_2 in the other. We find the median of the $(n_1 + n_2)$ observations and count the number of observations in each sample above this median. If the two populations have the same median, we would expect these numbers to be about $n_1/2$ and $n_2/2$. Deviations from this expectation are tested using χ^2 as a criterion. If we call the observed number of observations above the median in each sample m_1 and m_2 , we need only compute $\chi^2_{id t}$ from the tabulation

m_1	m_2	$(n_1 + n_2)/2$
$n_1 - m_1$	$n_2 - m_2$	$(n_1 + n_2)/2$
n_1	n_2	$n_1 + n_2$

Let us take as an example 30 subjects drawn randomly, of whom 10 are male and 20 female. They score in the following manner on a short attitude questionnaire

Men 12, 21, 28, 37, 38, 38, 39, 40, 42, 51

Women 12, 13, 15, 15, 21, 23, 24, 24, 25, 27, 30, 31, 32, 33, 34, 34, 36, 38, 43, 44

The median of the combined sample is 31.5—i.e., 15 observations are smaller, 15 larger. The chi square test is

	Men	Women	
Larger	7	8	15
Smaller	3	12	15
	10	20	30

Here $\chi^2_{id t} = [30(60)_1]/[10 \ 20 \ 15 \ 15] = 2.40$, with probability level about 12 percent. The t test of differences between means yields a significance level of about 8 percent ($t = 1.78$). We should conclude that we have no evidence of any difference between men and women on this questionnaire.

The example used to illustrate the run test can also be analyzed here, yielding a χ^2 table

7	5	12
5	7	12
12	12	24

$\chi^2_{.05, 1} = 0.67$, which is not significant even at the 50-percent level

When the combined sample contains an even number of observations with no tied observations at the median, there will be $(n_1 + n_2)/2$ observations larger than the median and the same number smaller than the median. When $(n_1 + n_2)$ is odd, the sample median is one of the observations and it is not determined in which cell that observation is to be entered. If we make the rule arbitrarily to include it in the class less than the median, the test in the long run will not be affected. The same arbitrary rule provides us with an approximate solution when ties occur at the median, that is, to call m_1 and m_2 the number of observations above the median, and $n_1 - m_1$ and $n_2 - m_2$ the number of observations less than or equal to the median. It is only approximate because our measurement is too crude to differentiate between observational values.

AN ANALOGUE OF THE ONE FACTOR ANALYSIS OF VARIANCE TEST³ (9)

This test is a straightforward extension of the median test in the same way as the simple F test is an extension of the t test. We go from testing the null hypothesis that two medians are equal to the null hypothesis that a number, k , of medians are equal, against the alternatives H , that at least one of the k medians is different from the others.

No information is available concerning the power efficiency of this test. Empirical studies comparing this test with the F test in normal populations have yielded quite good results.

Let us say that we have k different levels of the factor with n_i observations at the i th level. We find the median of all the observations, count the number of observations at each level above the median, the m_i 's, and enter these in the first row of a 2 by k table. In the second row, we enter the number $n_i - m_i$. The χ^2 test of independence of rows from columns is then a test of significance of differences among medians.

As an example, we shall analyze the following data from Edwards (2)

³Various other nonparametric analyses of specific experimental designs are discussed fully in Mood (9).

	Levels					
	1	2	3	4	5	
	13	7	12	10	13	
	9	4	11	12	6	
	8	4	4	9	14	
	7	1	9	7	12	
	8	10	5	15	13	
	6	7	10	14	10	
	6	5	2	10	8	
	7	9	8	17	4	
	6	5	3	14	9	
	10	8	6	12	11	
Mean \bar{X}	8	6	7	12	10	$\bar{X} = 8.6$
Med. (\tilde{X})	7.5	6	7	12	10.5	Grand med. (X) = 8.5

An F test of these data yields a highly significant result ($F_{4,45} = 6.52$). Our test is as follows:

1. The median of the 50 values is 8.5.
2. The 2 by 5 table is

	Level					
	1	2	3	4	5	
Above 8.5	3	2	4	9	7	25
Below 8.5	7	8	6	1	3	25
	10	10	10	10	10	50

$$3. \chi^2_{df=4} = \frac{(5-3)^2}{5} + \frac{(5-7)^2}{5} + \frac{(5-2)^2}{5} + \dots + \frac{(5-3)^2}{5} = 13.6.$$

4. This value, though not as significant as the value of F , still attains the 1-percent level. The test would not be expected to give

TABLE V

AFTER

Couple	BEFORE				AFTER			
	WIFE		HUSBAND		WIFE		HUSBAND	
	Rank	Score	Sign*	Rank	Score	Sign*	Rank	Score
1	16	47	-	12	51	-	14	61
2	4	81	+	16.5	43	-	4	94
3	8	62	+	13	50	-	7	84
4	12.5	49	-	9	59	+	16	56
5	5	79	+	3	91	+	2	101
6	12.5	49	-	15	47	-	18	49
7	14.5	48	-	7	69	+	12	76
8	3	86	+	5	83	+	6	88
9	10	61	-	16.5	43	-	10	78
10	1	110	+	2	100	+	1	109
11	6.5	72	+	11	52	-	11	77
12	18	36	-	18	41	-	8.5	79
13	2	97	+	1	109	+	3	98
14	14.5	48	-	10	53	-	15	58
15	9	64	+	6	71	+	8.5	79
16	11	57	-	9	55	+	13	71
17	6.5	72	+	4	85	+	5	92
18	17	43	-	14	49	-	17	54
Median		61.5			53.5			78.5

*A negative sign indicates that the observation is below the column median, a positive sign, above the median

as significant results, since the samples were drawn from normal populations. A further reason for our test's moderate lack of sensitivity is that we have used only an approximate value for our χ^2 test. Using the maximum likelihood value (9, p. 276) we obtain a $\chi^2 = 14.97$, which attains significance at the $\frac{1}{2}$ -percent level, much closer to the significance level of F . We would reject, at the 1-percent level, the hypothesis that the 5 samples were drawn from populations with the same median.

Tests of Relationship

This section will be devoted to a rather brief discussion of tests of association in a bivariate population. None of the methods in this section has been adequately investigated for power efficiency. They remain useful, however, especially when normality assumptions are intolerable.

One example (13) will be followed through the various tests discussed, and also through the following section on estimation. The data consist of scores on a thirty-item attitude-toward-children scale obtained by eighteen couples before and after an educational program. Five-point scoring was used, and the seventy-two scores obtained ranged from 36 to 110, the maximum possible range being 0 to 120. A high score indicates a "permissive" attitude, a low score a "rigid" attitude. The data are contained in Table V.

THE CONTINGENCY TEST (9) The first and simplest test is the test of whether the linear regression line of one variable on the other, fitted nonparametrically (9, p. 408), has a slope of zero. The test becomes just a test of linear independence in the 2×2 table formed by the X median and the Y median. Under the null hypothesis, one would expect about $n/4$ observations in each cell. Departures from this indicate association—i.e., high values of one variable—tend to be associated with high (low) values of the other. This is called positive (negative) association.

Four different tests of interest are possible on the data in our example. The first test will be, "Is there significant association between husbands' and wives' scores on the 'before' test?" The 2×2 table shows some positive association, but with a χ^2 of 2, is not significant at the 10-percent level. It should be noted that the cell frequencies are not large enough to make the chi-square test accurate.

		HUSBAND		
		<u>> med</u>	<u>< med</u>	
Wife	> med	6	3	9
	< med	3	6	9
		9	9	18

The same test on the "after" questionnaire can be made on the 2×2 table with the same results as before

		MEN		
		<u>> med</u>	<u>< med</u>	
Women	> med	6	3	9
	< med	3	6	9
		9	9	18

The other tests, women before and after and men before and after, show highly significant association, as one would expect. The phi coefficient can be used as an estimate or index of association, being $+0.33$ in these cases.

We should not expect this test to be very powerful, since only a few slight shifts in scores near the median could radically alter the conclusions drawn. Its ease, however, recommends it as a simple preliminary device to find highly related variables.

THE CORNER TEST (10) This test lays much more emphasis on the association observed between the two variables at the extreme values of each than does any similar test so far proposed. In application it is not quite so simple as the one above, but it would seem to be much more powerful. This emphasis on extreme values is a rather valuable contribution of the test, since one is often most interested in the characteristics of persons who deviate a great deal from the norm. This interest is often due to the feeling that minor extraneous factors may disturb the relationship between two variates where neither is

intense, but such factors should have relatively less effect when one or the other variate is present to a high degree.

For illustration we shall use the results of the first administration of the attitude scale mentioned previously. The scores are plotted in Figure 6.

A scatter diagram is used along with the medians, \tilde{y} and \tilde{x} . The four additional boundary lines are added as follows:

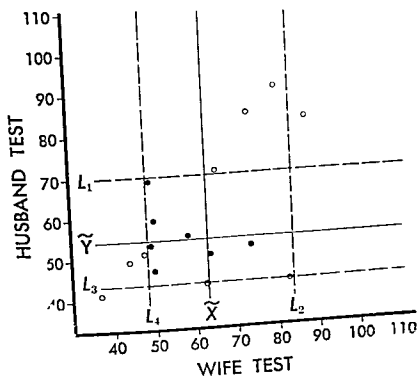


FIG. 6. An illustration of the use of the corner test on the data of Table V.

1. L_1 is drawn parallel to \tilde{y} and as close to \tilde{y} as possible under the condition that all points above L_1 are on one side of \tilde{x} . Call this the "on" side of L_1 . If L_1 were closer to \tilde{y} , a point to the left of \tilde{x} would be above it. If the first point on the "off" side is tied with a point on the "on" side, the boundary line is drawn through them.

2. Similarly, all points to the right of L_2 lie on one side of \tilde{y} , those below L_2 are on one side of \tilde{x} , and points to the left of L_4 lie on one side of \tilde{y} .

3 Points above L_1 are counted, and a plus or minus sign is attached according as the points are to the right or left of \bar{x} . In our illustration, this number is 6.

4 Points to the right of L_2 are counted and a plus or minus sign is attached according as they are above or below \bar{y} . In our illustration, this number is 3.

5 Points below L_3 are counted and a plus or minus sign is attached according as whether the points are to the left or right of \bar{x} . In our illustration, this is 1.5, counting only half of the tied point.

6 Points to the left of L_4 are counted and a plus or minus sign is attached according as they are below or above \bar{y} . In our illustration, this is 3.

7 These numbers are algebraically summed to get the test criterion (in our illustration, $r = 13.5$).

8 In the limiting case, as sample size becomes large, the 5-percent significance level is given by $|r| > 11$, and this criterion is adequate if sample size is larger than 10. For sample size of 11, not all points need to be outside the boundary, since "corner" points are counted twice. In our case, $r = 13.5$, so the null hypothesis is rejected at the 5 percent level.

It should be noted that, both in this test and the preceding one, the procedure cannot be followed without modification if there are ties either at one of the medians or on a boundary line, or if there is an odd number of observations.

When there is an odd number of observations, one observation will lie on each median. For example, the median of x values might be 71 and the median of y values 84, with the two observations being (71, 19) and (47, 84), and we could not tell whether either point was on the "on" or "off" side of the median. The ambiguity can be removed by substituting the observation (47, 19) for the two observations. This substitution does not bias the test in any way. In this way we have an even number of observations without affecting either median. It is possible, of course, that the odd observation lies on both medians. In our example it might be (71, 84). In this case we may neglect this observation, since it cannot affect the quadrant sums.

Ties not on a median are important only if they affect the position of one of the boundary lines, L . In Figure 6, the two observations (61, 43) and (81, 43) are tied on L_3 . Consequently, we do not know whether to count (61, 43) as outside or inside the boundary. The

authors of the test suggest that we should count it as $\frac{1}{2}$ outside. Had there been an observation (70, 43), we should have counted it as $\frac{1}{3}$. The general rule is to count the number of observations tied on the boundary as

$$\frac{\text{number on the "on" side of the median}}{1 + \text{number on the "off" side of the median}}$$

If in our example, 5 points had lain on L_3 to the left of \bar{x} and $^-$ to the right, the 5 points would contribute $\frac{5}{8}$ to the value of r .

RANK CORRELATION (5) This class of statistics is perhaps the oldest and best known of all nonparametric statistics. Originally conceived of as a shorthand method of estimating a product moment correlation, a conception which forced acceptance of many assumptions rarely satisfied in practice, plus several orders of mathematical approximation, they have come more and more to be considered as indices of association *per se*.

No assumption need be made about populations except that samples of observations can be arranged in rank order—i.e., that a simple ordering exists, and even where there are ties in the sample, reasonable approximations can be used.

The Spearman coefficient ρ (rho) is the earliest and best known of rank order statistics. It ranges between plus and minus one and is quite simple to compute. This, and the fact that it is closely related to the concordance coefficient, are the only reasons we have for including ρ , since Kendall's τ (5), to be discussed later, is in all other ways preferable. The computing formula is $\rho = 1 - [\sum d_i^2] / [(1/6)(n^3 - n)]$, where n is the number of objects ranked and d_i is the difference between the object's rank on one ranking and its rank on the other. For example, suppose four attitude items on anti-Semitism are ranked by two judges that is, how much anti-Semitism one would have if he endorsed each item. Suppose the rankings were

Item	1	2	3	4
Judge 1	3	1	2	4
Judge 2	3	2	4	1

Here $n = 4$, $\sum d_i^2 = (3 - 3)^2 + (1 - 2)^2 + (2 - 4)^2 + (4 - 1)^2 = 14$, and $\rho = 1 - [14] / [(1/6)(60)] = -0.40$.

When ties occur, the process becomes slightly more difficult. Ties

tend to raise, spuriously, the computed correlation between rankings, and a correction must be made. We can illustrate this case by supposing that Judge 1 had been unable to distinguish between Items 2 and 3. The rankings would have been

Item	1	2	3	4
Judge 1	1	2.5	2.5	4
Judge 2	3	2	4	1

and $\Sigma d^2 = 15.5$. In the denominator we would replace $(1/6)(n^3 - n)$ by $\sqrt{(1/6)(n^3 - n) - \frac{1}{2} \sqrt{(1/6)(n^3 - n)}} = \sqrt{95} = 9.75$, and ρ would become $1 - 15.5/9.75 = -0.41$. The correction for ties in general is that we divide by $\sqrt{(1/6)(n^3 - n) - T \sqrt{(1/6)(n^3 - n) - U}}$ instead of $(1/6)(n^3 - n)$ where $T = 1/12 [\Sigma (t^3 - t)]$ and t is the number of objects involved in one tied ranking, and $U = 1/12 [\Sigma (u^3 - u)]$, u being the number involved in the other. In the ranking 1, 3, 3, 3, 5, 6, 5, 6, 5, 9, 9, 9, for example, $T = (1/12) [(3^3 - 3) + (2^3 - 2) + (3^3 - 3)] = (1/12) [54] = 4.5$. If the other ranking were 3, 3, 3, 3, 3, 6, 7, 5, 7, 5, 9, 10, $U = (1/12) [(5^3 - 5) + (2^3 - 2)] = 10.5$. The correlation would be $1 - [(12.5)/\sqrt{(160.5)(154.5)}] = 0.921$.

In the attitude study example, the Spearman correlation between husbands' and wives' scores on the first administration is 0.586, on the second administration, 0.649. Couples seem to have become more alike following the training course, although we have no way of testing this. The correlation between the rankings of men before and after training is 0.765, the same coefficient for women is 0.804.

When n is larger than 9, we can test the hypothesis that $\rho = 0$, against the alternatives $\rho \neq 0$, by forming $t = \rho \sqrt{(n-2)/(1-\rho^2)}$ which is distributed as Student's t with $(n-2)$ degrees of freedom. For the correlation between husbands and wives on the first administration, $t = 0.586 \sqrt{16/[1 - (0.586)^2]} = 2.89$ with 16 degrees of freedom, which is not quite significant at the 1-percent level. For n smaller than 9, tables in Kendall (5) give significance levels.

When a number of rankings of the same objects exist, it is sometimes of interest to determine how much agreement there is among the rankings. For example, we might have a number of judges ranking a set of statements for pro-war feeling and we might wish to use average rankings to get a best estimate of some "true ranking," with which all judges more or less agree. We would want to be assured that there

exists some such ranking—i.e., that the judges were not ranking the statements at random. The concordance coefficient, W , is a measure of the amount of agreement among the judges. The coefficient $W = 1.0$ if all judges agree perfectly, zero if they disagree as much as possible.

The coefficient is given by the formula

$$W = 12S/[m^2 (n^3 - n)],$$

where there are n objects being ranked, m judges, and S is the sum of the squares of the rank sums for each object around a mean of $[m(n+1)]/2$. In our attitude-scale example, if we consider the two test scores of the husband and the two test scores of the wife as four rankings by "judges" of the couples, then the number of "objects" being ranked, the number of couples, is 18 and the number of "judges" is four. The mean rank assigned by each "judge" is 9.5, and hence the average rank sum is $4(9.5)$, or 38. Now the rank sum for the first couple is $(16 + 12 + 14 + 18) = 60$, for the second couple $(4 + 16 + 5 + 4 + 14) = 38.5$, for the third couple 44, and so on. The sum of squares, S , is then $(60 - 38)^2 + (38.5 - 38)^2 + (44 - 38)^2 + \dots$. In this case $n = 18$, $m = 4$, and $S = 5693.5$, so that $W = 5693.5/7752 = 0.734$. To test the null hypothesis $H_0: W = 0$ against the alternatives, $W > 0$, we can use the Snedecor (12) F test with $F = [(m-1)W]/(1-W) = 8.35$ with $[n-1-(2/m)]$ degrees of freedom in the numerator and $[(m-1)(n-1-[2/m])]$ degrees of freedom in the denominator, here 16.5 and 49.5. There is a highly significant amount of agreement and we can proceed with ordering the couples with respect to "permissiveness."

The average Spearman rank correlation between the 6 rankings is given by the formula $\rho_{s\bar{r}} = (mW - 1)/(m - 1)$, and in our example is 0.644.

KENDALL'S TAU (5) coefficient is devised for the same purpose as the rho coefficient mentioned earlier. Although somewhat more difficult to compute, it has the advantages of being generalizable into a partial correlation coefficient, and of having an almost normal distribution for samples as small as 9.

The computing formula is $\tau = 2S/n(n-1)$, where n is again the number of objects ranked and S is determined in this manner:

- (1) Order the objects according to one of the rankings
- (2) In the other ranking, for each object count the number of

objects to its right having a smaller rank and the number having a larger rank. Subtract the former from the latter.

(3) Sum the numbers thus obtained for all objects to obtain S . As a single example, consider the rankings 1, 4, 3, 5, 2, 6 and 2, 3, 4, 1, 5, 6. If we arrange the first in order, the order of the second becomes 2, 5, 4, 3, 1, 6. Start with 2. Since 3, 4, 5, and 6 are to its right and larger, and only 1 is smaller, the contribution is $4 - 1$ or "3." For 5, only one number to its right is larger, 6, and 3 are smaller, so that 5 contributes a "-2," the number 4 contributes a "-1," the object ranked 3 contributes "0," and the object ranked 1 contributes a "+1." These contributions added give $3 - 2 - 1 + 0 + 1$, or 1. This is the value of S . τ is $(2)(1)/(6)(5) = 1/15$.

In the case of ties, τ becomes

$$\frac{S}{\sqrt{0.5n(n-1)} - T} \frac{S}{\sqrt{0.5n(n-1)} - U}$$

For this coefficient $T = 0.5 \sum t(t-1)$ and $U = 0.5 \sum u(u-1)$, where t and u are used in the same way as in Spearman's coefficient. For example, in the ranking 1, 2, 4, 4, 4, 6, 5, 6, 5, 9, 9, 9, the correction is $0.5 [3 \cdot 2 + 2 \cdot 1 + 3 \cdot 2]$, or 7. If another judge had ranked the same objects 3, 3, 3, 4, 5, 4, 5, 7, 7, 7, 9, 10, S would be $7 + 7 + 7 + 5 + 5 + 2 + 2 + 2 + 1 = 38$, $\sqrt{0.5n(n-1)} - T$ would be $\sqrt{(0.5)(10)(9)} - 7$, $\sqrt{0.5n(n-1)} - 0.5U$ would be $\sqrt{(0.5)(10)(9)} - 7$, and $\tau = 38 / \sqrt{38} \sqrt{38} = 1$, as we might expect, since each ranking contained the same number of ties and was in the same order.

If we label the columns of Table V in order 1, 2, 3, and 4, we have $\tau_{12} = 0.457$, $\tau_{13} = 0.656$, $\tau_{14} = 0.477$, $\tau_{23} = 0.401$, $\tau_{24} = 0.559$, and $\tau_{34} = 0.493$.

When n is greater than 10, τ is about normally distributed with a mean of 0 and a variance of $[2/9] [(2n+5)/(n^2-n)]$. In our example, $\sigma_\tau^2 = [2/9] [41/306] = 0.029774$ and a standard error, σ_τ , of 0.1725. The normal deviate for $\tau_{12} = 0.457/0.1725 = 2.649$, which is significant just under the 1-percent level and does not differ much from the test of ρ for the same rankings.

To discuss partial correlation, let us consider the three rankings P , Q , and R .

P	1	2	3	4	5
Q	1	3	4	2	5
R	2	3	4	1	5

Let us make a table with an entry for each pair of objects. A plus sign is entered if the pair is in the order at the column head, a minus sign if it is reversed

	(12)	(13)	(14)	(15)	(23)	(24)	(25)	(34)	(35)	(45)
P	+	+	+	+	+	+	+	+	+	+
Q	+	+	+	+	-	-	+	+	+	+
R	-	-	-	+	+	+	+	+	+	+

The correlation between Q and R , partialling out P , is the phi coefficient in the following 2×2 tabulation

		RANKING R		
		<i>Pairs agreeing with P</i>	<i>Pairs disagreeing with P</i>	<i>Totals</i>
RANKING Q	Pairs agreeing with P	5	3	8
	Pairs disagreeing with P	2	0	2
	Totals	7	3	10

The phi coefficient is $-6/\sqrt{(2)(3)(7)(8)} = -0.327$. This is more easily determined from the formula

$$\tau_{QR.P} = [\tau_{QR} - \tau_{PQ}\tau_{PR}] / \sqrt{[1 - \tau_{PQ}^2][1 - \tau_{PR}^2]}$$

Here $\tau_{QR} = 0$, $\tau_{PQ} = 0.60$, $\tau_{PR} = 0.40$, and

$$\tau_{QR.P} = [0 - (0.60)(0.40)] / \sqrt{(0.64)(0.84)} = -0.327$$

The partial correlation can be interpreted as the correlation between the ratings Q and R when the factors on which P is rating are held constant. This interpretation holds whether the "judges" are actually judges, or tests, or any other instrument capable of arranging objects in rank order. Unfortunately, the distribution of partial taus is not known. In the short example just considered, we have some evidence, however, that when the factor on which the P ranking is based is held constant, Q and R are negatively related. P might be an intelligence test, Q a test of number ability, and R a test of musical

ability. In that case one would say that among persons having about the same intelligence, those having high number ability tend to have lower musical ability than those having low number ability.

An interesting use of this technique with our attitude-scale example would be to correlate the ranks of husbands and wives following the training period, partialling out the correlation due to the ranking of the couples before the training. To obtain a ranking of the couples before the training period, we can rank the sum of raw scores for husband and wife before training. We can call this ranking "*a*". The correlation between the women's scores after training and "*a*" is $r_{a3} = 0.669$, while $r_{a4} = 0.623$. The correlation between wives and husbands after training, partialling out the factors common to those present before training, is

$$r_{34 \cdot a} = [0.493 - (0.669)(0.623)] / \sqrt{(0.552)(0.612)} = 0.133,$$

indicating an extremely low relationship. This might be interpreted as meaning that the relationship between husband's and wife's scores after training is mainly due to factors present before training. If a partial correlation of 0.133 is significant for samples of size 18, we could also say that training tended to make husbands' scores and wives' scores more similar with respect to relative ranking in their respective groups.

ESTIMATION WITH ORDER STATISTICS

Percentiles (1, 9)

One of the most interesting results from order statistics is that the expected proportion of the population falling between two ordered observations is $1/(n+1)$, where n is the size of the sample. — *i.e.*, the sample is expected to divide the population into $(n+1)$ groups of equal size. Because of this, the percentile points of the sample are estimates of the percentile points of the population. For clarity we will define $X_{i:n}$ as the i th smallest observation in a sample of size n . For example, with a sample of 19, the fifth observation from the bottom, $X_{5:19}$ is an estimate of the lower quartile, the twenty-fifth percentile point, and in general, the i th observation from the bottom, $X_{i:n}$ is an estimate of the $100i/(n+1)$ percentile point. For a sample of 200, for example, the closest observation to the twenty-fifth percentile is the

fiftieth from the bottom, $X_{50|200}$ which estimates the 24.9th percentile.

If we require a 95-percent confidence interval for the 100th percentile, we look for the r th and the s th observation ($r < s$), $X_{r|n}$ and $X_{s|n}$, so that in approximately 95 percent of the samples of size n we draw, the p th percentile of the population will lie between $X_{r|n}$ and $X_{s|n}$. Any r and s that satisfy the equation

$$\sum_{i=r}^s \binom{n}{i} p^i (1-p)^{n-i} = 0.95$$

will determine such a $X_{r|n}$ and $X_{s|n}$, by the binomial theorem. In most cases we shall not be able to find an r and an s to satisfy the equation exactly, but we shall usually be able to approximate it rather closely. For a sample of size 10, we might wish to find a 95-percent upper confidence level for the tenth percentile—i.e., we might want a number which in 95 percent of the samples we could draw would be larger than the tenth percentile of the population. Here we choose r equal to zero and sum terms until we reach 0.95. The first term is $\binom{10}{0} (1/10)^0 (9/10)^{10}$, or 0.3483. This is the probability that the smallest sample observation is larger than the tenth percentile of the population. The second term is $\binom{10}{1} (1/10)^1 (9/10)^9 = 0.3871$, and the sum is 0.7354. This is the probability that the second smallest sample observation is larger than the tenth percentile of the population. Continuing, the third term is $\binom{10}{2} (1/10)^2 (9/10)^8 = 0.1936$, the sum of the first three terms is 0.929, and the sum of the first four terms is 0.9864. We see that the probability that the tenth percentile is less than the fourth smallest observation in a sample of 10 is about 0.93, the probability that it is less than the fifth smallest observation is about 0.99.

When we have large samples and wish to find confidence intervals for percentiles between the twentieth and eightieth, a normal approximation can be used.

A 95-percent confidence interval for the p th percentile is given by $X_{r,n} < p\text{th percentile} < X_{s,n}$, where r is $[p(n+1)/100] - [(1.96/100)\sqrt{np(100-p)}]$ and s is $[p(n+1)/100] + [(1.96/100)\sqrt{np(100-p)}]$. For example, let us find the approximate 95-percent confidence limits for the twentieth percentile in samples of 79. Here the point estimate is $X_{16,79}$, that is, the sixteenth observation from the bottom. Half the length of the interval is $[1.96/100]\sqrt{79(20)(80)} = 0.784\sqrt{79}$, or approxi-

mately 7. Thus, the 95-percent confidence interval for the twentieth percentile is $X_{9/79}$ to $X_{23/79}$. The interpretation is this: "In 95 percent of samples of size 79 drawn from a *continuous* population, the twentieth percentile of the population will lie between the ninth smallest and the twenty-third smallest observations."

Estimation of the Cumulative Distribution (1)

A method based on the same theory as the Smirnov test discussed earlier is available for estimating the cumulative distribution (ogive) of the population from that of the sample. In Figure 7, we have plotted

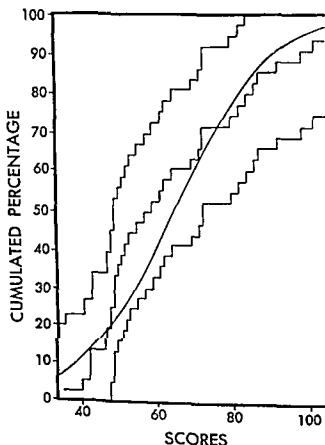


FIG 7 A comparison of the distribution of pretraining scores on the Stott-Berson attitude scale with a normal distribution having the same mean and variance

an example based on the 36 scores of men and women on the attitude scale before the training period. The center jagged line is the cumulative curve for the sample and the two outer lines are 90-percent confidence bounds for the population. That is, if we draw samples of size 36 from continuous populations a large number of times, we expect that 90 percent of the confidence intervals so computed will contain, *completely*, the population cumulative. The 90 percent limits are computed by finding $d_{0.90} = 1.22/\sqrt{N}$ where N is the sample size. In our example, $d_{0.90} = 0.20$. The upper confidence bound is parallel to the sample cumulative and shifted vertically 20 percentage points. The lower bound is parallel to the sample cumulative and shifted downward 20 percentage points.

If we wish to test goodness-of-fit of a sample to a theoretical distribution we need only plot the theoretical cumulative. If, at *any* point, it does not lie between the confidence bounds, we reject, at the level used in computing the bounds, the null hypothesis H_0 : the sample was drawn from a population with this distribution. The smooth curve is the cumulative distribution of a normal curve with mean of 68.22 and standard error of 20.125. Since it lies entirely within the bounds, we cannot reject the hypothesis that the sample was drawn from such a population.

CONCLUSION

The relative emphases of this chapter on tests of hypotheses and estimation should not be seen as an opinion of the relative importance of these two phases of statistical technique. Estimation of the magnitudes of differences in different populations must certainly be of prime importance for the scientific goal of prediction. But first, we must ascertain where differences exist. The point of view of this chapter has been that the social scientist searching for population differences must proceed with a maximally open mind, statistically as well as in his field of interest. After investigation has disclosed differences and use of the Smirnov or Marshall techniques has given information about the shape of the population distribution curves, powerful estimation procedures, based on the parameters which determine the distribution curve, can be used on *those* problems.

Investigations in the biological fields have shown that much of

their data is well fitted by normal theory. It remains an open question whether this is true in social science data. If it is true in our case, we are fortunate, if not, we shall at least know, and be able to act on our knowledge.

BIBLIOGRAPHY

- 1 Dixon W J, and Massey F J Jr *Introduction to statistical analysis* New York McGraw Hill, 1951
- 2 Edwards A L *Experimental design in psychological research* New York Rinehart, 1950
- 3 Fisher, R. A *Statistical methods for research workers*, 10th Ed Edinburgh Oliver and Boyd 1946
- 4 Geary, R C Testing for normality *Biometrika*, 1947 34, 209-242
- 5 Kendall M G *Rank correlation methods* London Griffin 1948
- 6 Mann H B, and Whitney, D R On a test of whether one of two random variables is stochastically larger than the other *Ann Math Statist*, 1947, 18, 50-60
- 7 Marshall A W A large sample test of the hypothesis that one of two random variables is stochastically larger than the other *J Amer Stat Assoc*, 1951, 46, 366-374
- 8 Massey, F J, Jr The Kolmogorov-Smirnov test for goodness of fit *J Amer Stat Assoc*, 1951 46, 68-78
- 9 Mood A M *Introduction to the theory of statistics* New York McGraw Hill 1950
- 10 Olmstead P S, and Tukey J W A A corner test for association *Ann Math Statist*, 1947, 18, 495-513
- 11 Pitman E J G Significance tests which may be applied to samples from any populations *J Royal Statist Soc Suppl* 1937 4, 119-130
- 12 Snedecor G W *Statistical methods* Ames Iowa State College Press 1946
- 13 Stott L H and Berson M P Some changes in attitudes resulting from a preparental education program *J Soc Psychol* 1951 34, 191-202

- 14 Swed, F S , and Eisenhart, C Tables for testing randomness of grouping in a sequence of alternatives *Ann Math Statist* , 1943, 14, 66-87
- 15 Vaart, H R van der Some remarks on the power function of Wilcoxon's test for the problem of two samples *Proc , Kon Ned Akad v Wetensch* , 1950, 53, No 4 146 172
- 16 Wald, A , and Wolfowitz, J On a test whether two samples are from the same population *Ann Math Statist* 1940 11 147 162
- 17 Westenberg, J Significance test for median and interquartile range in samples from continuous populations of any form *Proc , Kon Ned Akad v Wetensch* , 1948, 51, No 2 252 261
- 18 Wilcoxon, F Individual comparisons by ranking methods *Biometrical Bull* , 1945 1, 80 83

PART V

The Application of Research Findings

In the field of social research, more than in many fields, the research workers themselves are frequently concerned with the application of their findings and with problems of social action. There are two important differences between the social and the natural sciences in this respect. The usefulness of natural-science discoveries calls for little or no understanding of the principles involved by the user once the engineers have put the application into the form of a mechanical device. In social science, however, the findings having to do with interpersonal relations cannot be used without a real understanding of their meaning. The second difference is that application in social science means changing human behavior, and this is in itself part of the very subject matter of our field. Moreover, the social scientist becomes involved in the applications of his findings because he often maintains a close relationship

PART V

The Application of Research Findings

In the field of social research, more than in many fields, the research workers themselves are frequently concerned with the application of their findings and with problems of social action. There are two important differences between the social and the natural sciences in this respect. The usefulness of natural-science discoveries calls for little or no understanding of the principles involved by the user once the engineers have put the application into the form of a mechanical device. In social science, however, the findings having to do with interpersonal relations cannot be used without a real understanding of their meaning. The second difference is that application in social science means changing human behavior, and this is in itself part of the very subject matter of our field. Moreover, the social scientist becomes involved in the applications of his findings because he often maintains a close relationship

The Utilization of Social Science

Rensis Likert and Ronald Lippitt

The results of any piece of social scientific effort may have three kinds of usefulness. The findings and interpretations emerging from it may be of interest and value to *other social scientists* who will be asking such questions as "How does this piece of work give us further insight into individual and collective human behavior?

What new methods, hypotheses, and generalizations are presented in this study?" "How can I think and perform more intelligently as a social scientist because of the new facts and generalizations coming from this study?" Utilization by fellow scientists has as its goal the production of a greater body of scientifically valid knowledge.

The findings and interpretations may also be of value to *professional workers and citizen leaders* such as a business executive, a teacher, a community committee chairman, a labor leader, a government administrator, or a personnel trainer. These leaders have responsibility for public service and social action. Their goals are the improvement of policy making, planning, and acting in some area of social activity. In turning to social science, these leaders will be asking such questions as "What do these findings mean for the problems of planning and acting I am facing?" "Will anything

with the people, the groups, and the communities he studies, to ensure their cooperation in his research.

The following chapter deals with procedures and policies in the utilization of research results and the problems which arise in such application.

situation might be much more satisfying than it is now if it were different in certain respects. Perhaps as a result of talking with some other officers, the P T A president has become convinced that she should not be satisfied with 50 percent attendance at meetings. Perhaps certain clues have convinced the government administrator that there is a great deal of wasted effort in what is generally regarded as an important and effective program. Perhaps for some reason the chairman of a community council has the idea that all the competition that goes on between representatives of the various community groups who sit on his committee is not necessary. Perhaps the manager of a textile plant has read the reports of the experiments by Coch and French (9) and is wondering whether it is possible that the productivity of his plant which is regarded as quite good by industry standards could really shoot up in the same way. In each case this image or tentative question about potentiality stimulates need for further information about possibilities.

Some leaders and organizations have explicitly accepted the standard that a continuous effort to keep up with new discoveries and to try out new ideas is an imperative. For such persons and groups the utilization of science has become an important goal. This state of affairs is of course not yet common in regard to the social sciences but a growing number of individuals and organizations are explicitly establishing such a goal.

Even when there is motivation to turn to science for help this is just the beginning. Complex problems of research interpretation and application must also be solved. We shall examine these problems in two types of situations: (1) where there is a desire to apply scientific knowledge *discovered elsewhere* to the solution of a present problem and (2) where there is a desire to *apply research procedures directly* to help solve the present problem. We are making this distinction because it seems important for the analysis of the process of science utilization. In the first case there are questions as to whether and to what extent the research done elsewhere applies to the decisions and actions in question. Also there are questions of how the research from elsewhere gets communicated to the relevant actors in such a way that its practical value can be realistically assessed and acted on. In the second case we have the problems of whether the research is focused on major dimensions of the problem rather than on symptoms, whether the data collection activities

they found help me to do my job more intelligently or efficiently?

What have the scientists found that will help me right now?

The findings and interpretations may be of value also to the public — i.e. all citizens as individuals who are living in and adapting to the complex and changing twentieth century society. This utilization has the adult education goal of helping every man toward an understanding of the dynamics and potentialities of human activity as a means to the achievement of a more rational and satisfying personal and collective life.

In this chapter we are interested primarily in ways in which social practitioners and all citizens can utilize the resources of social psychology to improve personal insight, policy making, program planning and individual and group action. Other chapters of this volume have dealt with the standards and criteria by which scientists learn from one another.

The utilization of science will occur only if the person or group somehow becomes ready to look for and to use scientific resources in the solving of problems. This readiness and initiative seem usually to depend upon these three sources of motivation: (1) *problem sensitivity*, (2) an *image of potentiality*, (3) a general *experimental attitude* toward innovation.

Motivation for the use of scientific findings and methods often stems simply from the fact that the present state of affairs is unsatisfactory for someone. Perhaps the chairman of the P.T.A. program committee finds that attendance at meetings is showing a downward trend; the business executive discovers that the productivity of his plant is remaining steady or declining rather than showing improvement; the government agency is under attack from Congress to justify the way in which it has been spending funds; the solicitors in the Community Chest Drive are not collecting as much money as they have previously; a disruptive state of tension exists among ethnic groups in the community. John Doe feels he is not getting ahead in life. This type of sensitivity to a problem is frequently a reason why responsible leaders look outward for sources of help to get a deeper understanding of their problem situation and to find new principles and methods of functioning more effectively.

The *image of potentiality* is another very important source of initiative. Perhaps from their own imagination, or from observations of situations elsewhere, certain individuals have an idea of how a

generalizing from, or theorizing about his findings to provide helpful clues. It may sound somewhat paradoxical to state that one of the ways in which scientists can give their findings concrete significance for practitioners is to do adequate theorizing about the findings. But this is the case. The spelling out of abstract generalizations which emerge from the study of a specific situation provides one of the most helpful means of relating these insights to the analysis of other situations. The reports by Cartwright (6) on mass persuasion and Coch and French (9) on resistance to change are good examples of this. Cartwright's paper starts from the results of the studies of the selling of war bonds during World War II and the Coch and French paper starts from a study of the human factors involved in technological changeover in a textile plant. In both cases however, the authors moved to a level of theorizing about the phenomena they have studied which makes it possible for a wide range of practitioners to see how the generalizations apply to the analysis of their problems. This is possible because the concepts used to organize and interpret the data are concepts which are easily seen as relevant and important in a wide variety of situations.

Studies of Widely Distributed Phenomena and Populations

Studies such as those on the authoritarian personality, on autocratic and other kinds of leadership, on resistance to technological change, or on interpersonal relations between supervisors and workers have focused on aspects of behavior and social process which are important features of a wide variety of social problems in many types of social situations. Theoretical generalizations based on research dealing with widespread phenomena are likely to have relevance for a wide variety of practical applications. Social psychological studies which are focused on phenomena which occur infrequently in operating problems and situations are apt to yield generalizations of less widespread applicability. This does not mean that these studies are less basic as contributions to the developing science or to the solution of some specific problem.

There is another closely related way in which the scientist's approach to his research helps facilitate the process of research utilization. This is by the selection for study of situations and populations which are widely distributed in society such as industrial

have been accepted and understood by persons who it is hoped will utilize them and whether the research findings generalize to other problems and situations. We shall examine some of these problems of applying research and then review a number of illustrative cases of social science utilization.

USING KNOWLEDGE AND THEORY DERIVED FROM RESEARCH ELSEWHERE

As a person or organization turns to the scientific stockpile for help on a problem he is faced with a number of important questions about the applicability of research done in other settings. To what extent and in what ways is his situation comparable to those in which the research was done? Does a theoretical principle hold for his situation also? Is the way of approaching the problem all he can learn from the previous work or is there also something concerning the substantive content of the findings and generalizations?

Unfortunately many persons do not review these questions in looking for help from scientific resources hence they reject most scientific work and its implications because of certain manifest differences between the situations or populations on which the research was conducted and their own. Or they may uncritically accept all the findings and insights as relevant in their own situation and proceed unsuccessfully on this unrealistic assumption.

There is another significant question which must be asked. How can scientific knowledge about what causes what provide guidance in doing concrete thinking about what will happen if? Formulating *plans for action* on a scientific base often calls for more and different scientific information than the information needed to *understand why things are the way they are*. The sections which follow attempt to review some of the ways of thinking about these questions of applicability of scientific resources and some of the ways in which application can be facilitated.

The Need for Sufficient Theory

As the potential user reviews a piece of research done elsewhere he may or may not discover that the scientist has done enough

interpret the research findings and to observe the procedures being used. Peer reassurance is illustrated here by the fact that the visitors were able to talk about the feasibility and results of the project in their own nontechnical language with peers in the experimental community who were perceived as "the same kind of people we are." This experience encouraged the visiting leaders to try applying some of these principles in their own situations.

The staff of the Tavistock Institute of Human Relations has described a "budding-off conference" in which representatives of a factory where a major research project had been going on for some time invited representative visitor teams from several other industries to come to review what had been going on. Labor representatives talked to labor members of the host plant, engineers to engineers, and management to management as the first phase of the conference, before the social scientists were called on to help analyze what had been happening in the project. In many cases this type of communication is necessary to provide the motivation and insight needed for a "budding-off" of findings and methods to new situations.

We have mentioned previously that there is an additional problem of translating diagnostic insights about "why things are the way they are" into well-formulated hypotheses about "what to do about it" and how to test whether these alternative lines of action are correct and workable in a specific situation. This leads us to one of the most important processes of science utilization—the use of the scientist as a consultant. In a vast majority of cases, the effective carrying through of a process of utilization of research findings into integrated policy-making, planning, and operations requires active face-to-face interaction between a social scientist who serves as an interpreter and consultant and the key operating people involved. Such a scientific consultant is not primarily a producer of research. He is more of a social engineer and has a multiple role to perform. On the one hand he must become familiar enough with the operating problems so that he can help reformulate them to make possible a more scientific analysis of them. He must have a broad enough orientation to social research and theory so that he can bring the relevant research knowledge to bear on the analysis of the problem and the prediction of probable consequences of various lines of action. He must also be able to help set up procedures for measuring and assessing the consequences of new lines of action. Perhaps most important,

work groups, family units, parent-child relations, classrooms, etc. It should be clear, however, that certain basic problems can be most effectively investigated under laboratory conditions with populations of volunteers who are willing to collaborate in scientific experiments.

Channels of Communication

But even if the research setting, type of problem, and treatment of the data are advantageous to the providing of important and relevant insights to a wide variety of planners and actors, a third type of problem may nevertheless be present. There may be no actual communication of such relevant findings to potential consumers who need them. Effective communication must be established between relevant social scientific resources and the potential users of these resources. One help in this direction is the work of the social science 'middleman,' the science writer. A good social science interpreter is able to classify and synthesize research findings so that they are more clearly related to the problems posed by operating persons, and are related to a wide variety of problems so that many practitioners can find the material of relevance by reading such an overview and can learn where to find the sources of data. Examples of this are the books by Watson (39), Murphy (35), and Marrow (33). Such overviews may also be presented as specially invited papers by scientists at the professional meetings of practitioners (25). Overviews may frequently have the stimulus value of creating the *image of potentiality* referred to above, or they may provide someone who is *sensitive to a problem* with a direction in which to seek for help.

In the field of social psychological research, the implications of the findings for what to do in a specific social situation are often complex and likely to be tied up with fears and uncertainties about the consequences of the findings. Therefore, more adequate and intensive methods of communication are required to stimulate an understanding and acceptance of research findings and their implications than in the physical sciences. For this reason, the processes of *demonstration* and of reassurance by peers are often of great importance. For example, a very important community project on the solving of intergroup tensions sponsored an "open house" workshop for leaders from other communities to come and help review and

diagnostically some of the differences between "last year" and "this year" as a money-raising situation.

It is possible that all these hurdles might have been overcome had the consultant started out with the objective of creating a need for help, rather than assuming that this already existed. Or he might have been successful in suggesting a policy of "trying out" a modified approach with one part of the population.

COMPLACENCY SHOCK. As we have indicated previously, it is often difficult for the individual or the group to perceive and accept the fact that the operating situation needs improvement. The following case illustrates some of the elements involved in this type of situation where there is no strong problem sensitivity or image of potentiality:

The Mutual Security Agency has invited to this country numerous productivity teams from the various industries of the Marshall Plan countries. These teams of twelve are delegated to make a study of productivity in several American plants in order to find ways of improving the productivity in their own industry. As might be expected, there is in many cases a strong tendency to see any higher productivity in the American plants as due to special advantages not possible abroad, such as superior equipment or raw materials. In order to facilitate comparative analysis of their own plants and the American plants, the members of the team from one industry were asked to make a special pretrip study of productivity in their own factories by going around as a team visiting the various factories from which they came. They made a careful record of productivity data and of the various manufacturing procedures in their plants. The first plant they visited in the United States had about three times greater productivity per worker than their own plants. It was quite easy, however, for them to point out a variety of factors of superior equipment which they felt could account for the differences. They were then taken to a small plant comparable in size and equipment to a typical plant in their own country. After looking at this situation, they agreed that it was comparable to their own and therefore that a comparison of the figures on productivity should be valid. The analysis of productivity again indicated that the productivity per worker was two to three times as much as in their own factories. In the face of this situation, their complacency was

such a scientist consultant usually finds that he needs to take the role of a trainer of his operating co workers in the scientific attitude or scientific approach to thinking about the new operating problems which occur daily

Illustrations of Research Utilization

Now that we have reviewed briefly some of the problems and possibilities of motivation and communication involved in applying social science knowledge from other settings to specific operating problems let us look more closely at some examples of attempts to apply social science. In a later section we shall summarize what seem to us some of the important general principles that underlie these illustrations

The executive committee of a Community Chest asked a social psychologist to help them train their fund solicitors. After looking into the past practices of the solicitation procedure, the consultant recommended a number of modifications derived from research and theory. For example he suggested that solicitors participate more actively in the process of setting quota objectives so that they would feel psychologically more committed to their objectives. This was an application from several research findings that persons who participate in a group decision are more likely to carry through the commitments of this decision than are persons who receive assignments or exhortation without participation. Also the plan included more careful consideration of matching the solicitor to his targets in terms of existing status and group relationships. Research on social influence has indicated the importance of prestige and reference group membership in exerting influence. Another application of reference group theory was a recommendation to develop a clearer rationale for expected size of contribution in terms of specific subpopulations to which each giver could see himself as belonging.

The executive committee rejected all such ideas with a reaction that we made our quota last year by the previous methods so we'd better use them again. Psychologically they were unable to (1) accept the validity of the data from elsewhere as relevant (2) accept as realistic a goal of doing better than we have before (3) see

was considerable consensus in reformulating many of the operating problems. During this phase, the practitioners were primarily listeners although they asked questions of clarification.

During the third phase of the conference the whole group worked together on three activities: (a) redefining the nature of the basic operating problems in adult education, (b) formulating some general principles for the improvement of practice, and (c) identifying certain areas of needed research which should be conducted in adult education settings to test the value of certain theories developed in other settings or to open up new fields of knowledge which had not been explored in other settings.

The participants were in general agreement that this type of communication situation was very valuable and successful in clarifying the relevance of social science research to various operating problems and in reformulating operating problems so that they could be related to basic scientific findings and theory. Several problems of making this kind of interaction a successful communication process were identified. It was clear, for example, that the discussion coordinator needed, if possible, to be sensitive to the frame of reference of both the social scientists and the adult educators in order to identify and help clarify points of noncommunication and to help both groups find rewards in this type of interaction situation. It seemed to be particularly important to help members of both groups clarify their roles as listeners and as actors in various phases of the communication process. The coordinator also had to help give the conference a continuous movement toward applying generalizations from research to operating situations and formulating the major problems for further research.

A RESEARCH REVIEW CONFERENCE Our second illustration of research communication is a somewhat different type of research utilization conference. This was a one day meeting at which a heterogeneous group of fifty community social welfare leaders met to listen to a review of research on leadership and group dynamics.

During the first half hour the conference leader discussed with the community leaders ways in which they might plan to get as much as possible out of the research which would be presented. It was agreed that they would listen with an active interest in making notes on points in the research review at which each of them got glimmers of possible relevance to situations or operating problems about

shattered. They became eager to learn how to apply the discoveries and innovations of the American plants in their own. This illustration seems to be typical of many situations in which an individual or an organization has overtly expressed a desire to improve the present situation but is unable to accept various ideas for improvement as relevant for themselves until they can see that the improvements have been successfully executed by someone in their own league. When this factor is accepted, the data which have been discovered elsewhere are accepted as relevant and applicable to the analysis and evaluation of their own situation.

A RESEARCH APPLICATION CONFERENCE Let us turn now to several illustrations of attempts to communicate research findings in a way that would clarify their relevance to social practice. In the first example, a group of adult education practitioners decided they would like to explore what help they could get in thinking about their professional operating problems from interaction with a selected group of social scientists. They organized a two-day conference which had the following design.

During the first phase of the conference, the practitioners sat around the table taking a census of what they regarded as critical professional problems, attempting to arrive at some consensus concerning the most important problems as they saw them. They attempted to formulate the nature of these problems and the causes as they understood them. During this phase, the five or six social scientists representing various disciplines from anthropology to psychology had the job of listening and keeping notes on their interpretations of the nature and the causes of the problems, making cross references to any relevant social science research and theorizing which might be helpful in shedding light on the various problems which were being explored.

During the second phase of the conference, the social scientists took the center of the stage to interact with one another. They shared and integrated their observations about the crucial dimensions of the operating problems which had been discussed, the way in which these problems might be related to deeper underlying problems on which research had been done, and the kind of generalizations which might be tentatively advanced as guide lines for clarifying the nature of the operating problems. A wide variety of researches and research-derived generalizations were brought to bear, and there

about a problem. They also felt that the research reviewer had not had enough opportunity to get insight into the complexities of some of the operating problems which were being talked about so that he could sense more thoroughly the relevance or lack of relevance of some of the research findings and theory which he had been reporting. It was also observed that the participants in the discussion became quite motivated to read more social psychological literature, but the research reviewer had not been adequately prepared to suggest available types of reading which might provide good follow up.

FOCUSING ON A SPECIFIC OPERATING PROBLEM During World War II, one of the present writers had responsibility for a training program in which it was necessary to bring the disciplines of anthropology, social psychology, psychiatry, economics, political science, journalism, and geography to bear on the solution of a specific type of operating problem. To integrate this diverse spectrum of information into a focused set of operating knowledges, attitudes and skills, it was necessary to train each group of operating personnel (about a dozen in each subgroup) to operate as a learning team in utilizing information from many different scientist experts. For example, before a political science specialist arrived, the operating group worked together for a class period on the preparation of a group interview schedule of questions which would be covered and cross checked in conversational style during the group discussion with the specialist. The findings of the group discussion would be summarized after the specialist left, and hypotheses would be formulated for further exploration and cross-checking with other scientists.

In some cases the trainees needed to convert the information they acquired into specific interpersonal skills. For example, they needed to get from the anthropologists the kind of information which would actually help them behave appropriately in cross-cultural contacts with representatives of a culture very different from their own. They found that descriptions of appropriate behavior acquired by questioning the anthropologists were not easy to translate into actual behavior. They found it necessary and effective, therefore, to set up role playing to deal with specific cross-cultural contact situations, with the social scientist giving demonstrations and providing at the-elbow supervision in appropriate behavior patterns.

Another procedure for cross-disciplinary integration used with

which they were concerned. At the end of the research review they would convene in committees of six or seven to share their thoughts and observations in order to formulate (1) questions of clarification to the social scientist, and (2) tentative generalizations which they would draw from the research review concerning operating problems. These they would test by getting the reaction of the research scientist to the validity of their generalizations and the possible extension of the generalizations.

The research review which took about an hour and a half, integrated a great variety of studies under several research topic headings such as Leadership, Communication, Decision making, Participation, etc. Under each heading a variety of empirical results were presented and a number of theoretical generalizations were formulated. During the next hour, all the subcommittees held very active discussions, each committee having a recorder/reporter. The following period of interaction between the group reporters and the social scientists was also very active, although rather frustrating at times for the social scientist. The subcommittees had been very inventive and creative in applying and extending various research generalizations and he found himself in a position of challenging some of the applications which were being suggested. He felt that they needed considerable testing before they could be applied. He was active in suggesting boundaries to generalizations and dangers of oversimplifying the complexities of operating situations by thinking only in terms of the interaction of two or three variables rather than of many additional ones which might be just as important and which had not been explored in the research which he had been reporting.

As a whole, the conference seemed to strike a balance between providing new insights into the analysis of operating problems and training in "the scientific attitude" of thinking about some of the boundaries of generalizing from one situation and set of variables to another. The participants in the discussion indicated a high degree of satisfaction with the results. The research reviewer and the conference leader felt that there were a number of weaknesses which could be improved upon next time. They felt the whole problem of adequate research utilization had not been grasped because no single illustration had been carried all the way through to an examination of the problems of acting on new ideas and insights.

of social science utilization. For example, after agreement was reached concerning the over-all outline of content for a specific issue of the periodical, some of the social scientist members of the Committee undertook to prepare certain articles, and other members of the committee and staff with specialized writing skills undertook to produce other units, after a briefing in the committee meetings about the main themes of the content. Tension became somewhat high when the writers told the social scientists that their productions could not be accepted without rewriting and the social scientists told the writers that their contributions oversimplified the basic data which the article had been designed to interpret. Out of these first efforts have come a number of interesting production patterns. For example, certain members of the periodical staff have now acquired skills in briefing social scientists about what kind of content is needed in rough draft form to provide the basis for a writing job by a professional writer, whose effort will be reviewed by one or more social scientists. Another successful pattern of production has been to set up a briefing session during which one or more social scientists think out loud about what the content and sequence of a particular article might be while the professional communicator takes notes, asks questions of clarification, checks whether he is getting the ideas by rehearsing them out loud, and then writes up a first draft of the article for review. Out of this kind of interaction has grown a very real mutual respect on the part of both parties and a much deeper appreciation of the problems involved in an adequate process of transition from research findings and scientific theory to communications of functional significance to laymen who have operating jobs of leadership to perform.

DIRECT CONSULTATION ON A SOLUTION OF AN OPERATING PROBLEM
In one city, the school population had grown rapidly and there was need for an expansion of school facilities. Arrangements had been made for a local election in which voters would have an opportunity to vote for or against new schools and also for or against a bond issue to finance these schools.

Although the local Parent-Teacher Association was thoroughly familiar with the reasons for the need for the new school and was particularly anxious that the outcome of the vote be favorable, they felt that the vote should represent the desire of the entire community. Therefore, they decided to start a campaign to inform people about

these trainees was to give them a specific operational problem to solve in which their first job was to make decisions as to what kind of information they needed and from what type of social scientists and to formulate a procedure for getting this information from such scientific personnel. The scientists were then called in and the group had to make good use of them and integrate the material into a set of operating decisions and actions. It was the judgment of the three leaders of this training program that the focus on the solution of specific operating problems was an indispensable requirement for the successful integration and use of the resources of the various social sciences. One of the problems of such an approach was that of training the social scientists to enter into this group interview procedure in which they were asked to focus their thinking and background knowledge on certain functional questions rather than to present prepared lectures organizing the substantive content of their specialty around more general or abstract topic headings. Most of them proved quite ready to make the adjustment and were able to contribute in a very fruitful way.

PRODUCING AN APPLIED SOCIAL SCIENCE PERIODICAL Another example of the interpretation of the social sciences is illustrated in the production of the monthly periodical *Adult Leadership*. The professional association which has responsibility for the production of this periodical has conceived its main function as communicating to lay leaders principles and techniques for performing their job of community and organizational leadership. It was felt that this objective called for drawing on the resources of the social sciences in order to produce valid content for such a periodical. On the other hand, it was recognized that most social scientists would not have the time or the skill to communicate clearly through the type of writing demanded for such a popular periodical. Therefore an Operating Committee was established which was composed of representatives from the social sciences, from training methodology, from magazine publication and art layout, from public relations, and from the operating field of leadership practice. This group was given the major responsibility for thrashing out with the small periodical staff the major policies and procedures for the content and production of the periodical.

In the early stages of production fervor, a number of things happened which helped to put the microscope on certain problems

agency to help understand and apply the local results. A social science team was asked to undertake a program of research to evaluate the effectiveness of the foreign assistance program of one of the government agencies. It became clear that in order to understand and use the type of results which would be coming from the study it would be essential that the program planning staff and key administrators of the agency understand some of the dynamics underlying attitude formation, group action, and resistance to change. Before the study was completed, therefore, the key members of the staff were given an opportunity on a volunteer basis to participate in a seminar on changing attitudes and behavior.

The seminar was restricted to 15 participants who would commit themselves to attend all the sessions in spite of the operating pressures of their daily jobs. The one session a week was held several blocks away from the office building. The ground rules of the seminar were that research findings and general theoretical principles which might be relevant to the operating problems of the various members of the group would be presented for discussion but the discussion would not wander off into the specific discussion of any of the operating problems of the members of the group until the final sessions of the seminar. The last half hour of each two hour period however was set aside for brain storming. In this period each member was asked to free associate any connections he had made or was able to make between the things that had been reported and discussed in the session and his own operating situation. The seminar leaders kept a record of these free association connections for later review and discussion when there was a specific focus on the review and analysis of operating problems of the agency. This review also provided a transition to looking at the research findings of the evaluation study of their own program. It was the feeling of the research team that these background theory and research report sessions provided the kind of sensitivity and perspective which helped this group of operating persons to be able to come to grips effectively and objectively with the findings of the research conducted in their own situation.

THE ANALYSIS OF THE SITUATION One of the most difficult and important problems for the social scientist who is serving as a consultant is that of getting an accurate picture quickly of just what the operating problem is so that he may be able to select and

the facts to be voted upon and to get out the vote. In planning the campaign the head of the local P T A sought the advice of a social psychologist on what the P T A should do in order to be certain that the voters acted when the time came to vote.

This social psychologist had had research experience during the war on the problem of why people did or did not buy war bonds. He knew that the studies of the several war bond campaigns had shown that the mass media were useful in helping people to understand the reasons why it was important for them to buy bonds. But he also knew that a major finding of the studies of the campaigns was that people were likely to act by buying a bond only when they were personally asked to do so.

In advising the local P T A he drew upon his knowledge of these and related results. He urged them to plan a campaign utilizing both the mass media and personal solicitation but with emphasis on the latter. He suggested that they ask the local newspaper and radio stations to carry as much information as possible on the facts of the situation and to repeat this information in somewhat different stories on different days so as to increase the probability that people would become aware of the election and the issues involved. But he emphasized in talking with the P T A that if they wished to motivate people to participate in voting it would be very important to have each household called upon by a volunteer from the P T A to encourage every eligible voter in the household to vote. Acting in a vigorous manner upon his advice the P T A organized an information program about the need for additional schools and the cost of these additions which was widely disseminated through the cooperation of the local mass media. Most important of all however the local P T A organized a campaign in which every household throughout the city was called upon. In these personal calls on individual households neighbors of the person called upon gave him facts about the situation and urged him to be sure to vote in the forthcoming election. The effect of this campaign with its house to house solicitation was a large vote in the election and one which was overwhelmingly in favor of the additional schools and of the bond issue to finance them.

AN IN SERVICE SEMINAR A final illustration of methods of research interpretation is a combination of scientific discoveries from other situations and the findings of research conducted within an

of expectations, nonparticipation in decision making or noncommunication of decisions which had been made

To check further whether these were the most relevant phenomena, the consultant recruited several volunteers from each role to recreate some typical interaction episodes using role playing techniques. From his observations of these situations the consultant felt fairly secure about some of the major variables which were involved and was able to summarize some of the relevant research from other social hierarchical situations, and to make some general interpretations about the problems involved. Also he suggested from experience elsewhere that key personnel be actively involved in an actual research project in which they would participate in a more thorough diagnosis of the dynamics of their own problem situation. This step takes us into the next section of this chapter on the utilization of research done within the operating situation. The main point which this brief example illustrates is that the consultant can use a variety of techniques to get at least crude data on the particular operating problem so that he can more appropriately mobilize the scientific resources of research and theorizing done elsewhere which may be relevant in stimulating insights about the present problem.

Interpretative Review of Case Illustrations

IMPORTANT VARIATIONS IN THE APPLICATION SITUATION One of the important facts we can note from the variety of illustrations above is that the psychological readiness and ability of the participants to use social science and social scientists differs greatly from situation to situation. We have noted that in some cases the operating personnel have taken initiative to seek new knowledge and ways of applying it, whereas in other cases this sensitivity to potentialities does not exist. Also where this desire to seek new information does exist, in some situations it is a general attitude of curiosity about possibilities of application (such as in the research application conference and the research review conference) rather than a search for help on a specific problem (such as the school bond campaign or the improvement of the morale in the hospital). The task of the social scientist is somewhat different when the practitioners ask him,

interpret relevant research results and theoretical generalizations developed elsewhere. The conscientious research scientist is quite likely to take the point of view that this is an impossible task without an additional comprehensive research diagnosis based on measurement of the present situation. Frequently this is correct and this will be the focus of the second section of the chapter. Frequently however this approach is not feasible and would require quite an impractical delay in the decision making and planning demands. Therefore the social scientist consultant is frequently faced with the necessity of getting the picture as quickly and as objectively as possible. The following case illustrates two techniques of tackling this problem.

The social scientist was asked to consult with the administration of a large hospital on any relevant knowledge there might be which would be helpful in reducing the intergroup tensions among the nurses, doctors, and attendants in such a way that there would be improved care of the patients. During one afternoon the consultant conducted three problem census group sessions. In one session he met with a group of medical personnel, in another with a group of nursing personnel, and in the third with a group of attendants. In each session he shared his problem of wanting to get valid data on the problems each subpopulation felt it had of cooperation with the other subpopulations in the interests of serving the patient. He selected three persons at random in each group, explaining that while he interviewed them in front of the rest of the group he would like to have all the other members jot down whether they agreed with or disagreed with the responses of the interviewee and what modifications and additions they would make in their own response. The data would be reported orally if there was time but would be handed in if there was not sufficient time.

The interviews were then conducted with a preplanned set of questions and proceeded slowly enough so that the rest of the group could formulate the variations in their responses. Most of them were very active in doing this. The interviews and supplementary data indicated that most of the tensions and low morale stemmed from interaction situations which occurred in the daily routines of hospital life. It seemed to the consultant from these reports that probably a good deal of the difficulty had to do with noncommunication.

of expectations, nonparticipation in decision making, or noncommunication of decisions which had been made

To check further whether these were the most relevant phenomena, the consultant recruited several volunteers from each role to recreate some typical interaction episodes using role playing techniques. From his observations of these situations the consultant felt fairly secure about some of the major variables which were involved and was able to summarize some of the relevant research from other social hierarchical situations, and to make some general interpretations about the problems involved. Also he suggested from experience elsewhere that key personnel be actively involved in an actual research project in which they would participate in a more thorough diagnosis of the dynamics of their own problem situation. This step takes us into the next section of this chapter on the utilization of research done within the operating situation. The main point which this brief example illustrates is that the consultant can use a variety of techniques to get at least crude data on the particular operating problem so that he can more appropriately mobilize the scientific resources of research and theorizing done elsewhere which may be relevant in stimulating insights about the present problem.

Interpretative Review of Case Illustrations

IMPORTANT VARIATIONS IN THE APPLICATION SITUATION One of the important facts we can note from the variety of illustrations above is that the psychological readiness and ability of the participants to use social science and social scientists differs greatly from situation to situation. We have noted that in some cases the operating personnel have taken initiative to seek new knowledge and ways of applying it, whereas in other cases this sensitivity to potentialities does not exist. Also where this desire to seek new information does exist, in some situations it is a general attitude of curiosity about possibilities of application (such as in the research application conference and the research review conference) rather than a search for help on a specific problem (such as the school bond campaign or the improvement of the morale in the hospital). The task of the social scientist is somewhat different when the practitioners ask him,

"Tell us in general what's going on and what new things are being discovered in your field," from what it is when he is asked, "Do you know of any knowledge that will help us in solving this problem?" We can note another difference in our case illustrations between those situations in which the social scientist is primarily an information-giver and situations in which he is an active consultant in guiding the process of research application and doing an active job of interpreting scientific method and the generalization of scientific findings from one situation to another.

Our case illustrations emphasize the importance of this more active role for the scientist in applying scientific knowledge and theory in an appropriate and effective manner. This does not mean that many important applications have not also emerged from a practitioner's reading some book of social-science interpretation or hearing some research-interpretation speech. Such illustrations are not very likely to come to our attention. It is our hunch, however, that the complexities of social-science research utilization are so great that in most cases it requires the very active teamwork of motivated practitioners and well-oriented social scientists to bring about intelligent application of available scientific knowledge and theory.

THE PROCESS OF RESEARCH UTILIZATION. A review of the foregoing case illustrations from a different point of view reveals a number of things that rather typically need to happen if successful application of scientific research and theory is to result. Even though the situations did differ greatly from one another, the following common elements seem to be necessary or at least highly desirable:

(1) There needed to be the motivation to seek and use scientific resources. In situations where this motivation did not exist, it needed to be stimulated by demonstrations of potentiality, by complacency shock, and other approaches. This seemed to be the first step before further progress could be made.

(2) Then an active process of redefining and reformulating operating problems was required so that the relevance of scientific research done elsewhere could be perceived. In the research-application conference the team of social scientists did an active job of first listening to the practitioners' statements of their problems and then attempting to break these problems down and reformulate them

for scientific analysis. In the case of the research-review conference and of the administrative in-service seminar the practitioners took the initiative in attempting to relate their operating problems, as they saw them, to various research problems which had been worked on.

(3) In nearly all of the case illustrations, we saw that the social scientist, in order to try to be helpful, had to do an active job of getting oriented to the action problem as the practitioner saw it in order that he might do an intelligent job of selecting appropriate scientific resources for application to this particular situation. He had to be a skilled listener and interrogator, and in one case he had the persons involved recreate some of the typical social interactions of the problem situation in order to get an appropriate orientation. In many cases of social-science application an even more thorough diagnostic research job is necessary to get the facts about the operating situation. This type of research is discussed in the next section of this chapter, where we deal with the utilization of research conducted within the setting.

(4) It usually is helpful for the social scientist to communicate a general orientation or "way of looking at behavioral dynamics" as a framework within which to interpret specific data. This certainly suggests that some types of general education in the behavioral sciences is needed as a background for effective application of research knowledge to specific problems.

(5) In addition to reformulating operating problems in terms of relevant scientific findings or variables, an active job of thinking about generalization and applicability of research knowledge and theory is required. This requires that the social scientist, as a consultant, help the applier gain an understanding of the methodology of research application by facing such questions as comparability of populations, comparability of situational dynamics, extrapolation of theoretical generalizations to different situations, and experimental-mindedness in trying new solutions.

(6) One other important aspect of research application has been mentioned. The scientific knowledge which is related to a given operating situation or a problem may be of several types. For example, it may give insight into "why things are the way they are," it may give perspective on "ways in which the situation could be different," or it may give information about "how we can go about

changing the situation. In all cases, successful application requires that the applicer interpret, plan, and execute specific steps of action in his own situation. This requires creative and realistic thinking about 'what would happen if'. Helping formulate correct hypotheses and plans at this point is a challenging task. Even though the directions for application may be well understood, it may be necessary to acquire new technical or human relations skills to execute the desired steps.

RESEARCH METHODS APPLIED TO PROBLEMS OF ORGANIZATIONS

Social science research is being done increasingly in real life settings and on problems of major social significance. This trend is likely to continue as the methodologies and substantive findings of the social sciences become more extensive and more able to shed important light on complex problems. But if an organization is to derive full benefit from any research done on its problems, experience suggests that the work must be organized, planned, and conducted in accordance with certain patterns and principles. The rest of this chapter is devoted to a consideration of these patterns and principles. It is written primarily from the standpoint of a research staff which conducts sample surveys in, or for, an operating organization but the principles stated are believed to be generally applicable.

Establishing a Working Relationship

CREATING A COOPERATIVE ATMOSPHERE An important problem in conducting research in an organization concerns the attitude of the research staff toward the organization and its personnel. Since organizations vary greatly in their traditions and values, a researcher may wish, at times, to conduct research in an organization whose values and convictions differ in important ways from his own. In such circumstances, it is important that he act in a way that makes clear that he respects the values of the organization and of its members even though he may not completely agree with them. In planning and conducting the research and in reporting the results

to the organization, he will obtain better cooperation when he displays a sensitive appreciation of the values of the organization.

This does not mean that the researcher, himself, need feel compelled to accept and be guided personally by the values of the organization. He can maintain his own values and his personal integrity provided, of course, that they are not in direct conflict with those of the organization. When the researcher's values and those of the organization are in direct conflict, he usually is wise in not conducting research for the organization.

The major purpose from an organization's standpoint for conducting research is to improve its operation. But improvement always requires change. This problem is not new. Organizations are continuously undergoing some change. All that research does is to increase the amount and magnitude of change. All the problems involved in changing the activities of an organization consequently, are present when attempts are made to apply research findings.

It is common experience that orders by themselves are seldom sufficient to produce effective change in an organization and its functioning. Other procedures, including those which make some use of participation, are usually required. The persons who need to seek the participation or cooperation of the others are those persons who possess information as to what changes might bring improvement. When this information is based on research it is the researchers consequently, who are primarily faced with the problem of obtaining the participation and cooperation of the others. If the research results are to be applied successfully. Moreover, applying new ideas requires not only a knowledge of the new idea but also a full understanding of the present operation. The research staff consequently, faces the problem of obtaining participation not only to facilitate cooperation in bringing about desirable changes but also to be sure that the changes sought represent the best available thinking based both upon past experience and current research findings.

Much of the rest of this chapter will be devoted to a consideration of how to obtain this kind of participation. Cooperation in seeking and achieving change grows out of honest participation with full recognition and appreciation of the important ideas that the many kinds of people involved can contribute. Cooperation is not created by manipulation—at least not for long.

AVOIDING RESISTANCE When an organization is contemplating having research undertaken for it, there may be some important persons in influential positions who view the proposed research with some reservations. It is well not to proceed with the research until these criticisms and reservations have been examined fully and in relation to the advantages and disadvantages of undertaking the research. Their resistance will manifest itself sooner or later, and it usually is better to have it out in the open and fully faced early in the proceedings. Often, if this is done candidly and unemotionally, these persons will become more and more involved in the research and increasingly favorable in their attitude toward it. If resistance is ignored or brushed aside it is likely to result in efforts to stop the research if difficulties are encountered, or it may result in attempts to block applications of the research.

CREATING REALISTIC EXPECTATIONS Just as some persons are unduly skeptical as to the probable value of the results that research will yield, others are unduly optimistic. This latter group tends to have unrealistic expectations as to what research can do for them or for the organization. This can result in serious difficulties for there are aspirations that even the best research cannot possibly achieve. If people within an organization maintain these unrealistic expectations, they are bound to be disappointed with the results obtained from any study, no matter how good these results are. Such disappointment may lead them to reject the idea of using research in the future.

In order to avoid the disappointment which occurs when unreal expectations exist, it is important to create expectations that are modest in relation to the probable contributions from research. The creation of these moderate expectations is best done during the planning stages of the research project. This can be done while the problem to be studied and the probable character of the results to be obtained are discussed. When expectations are modest, the value of the research results are likely to be greater than anticipated. This encourages the further use of research.

Organization of Research Relationships

AN INTERNAL STAFF VS AN OUTSIDE RESEARCH ORGANIZATION
There is neither a simple nor a universal answer to the question of

whether research should be conducted by a unit in the organization or by an outside research staff. The answer depends upon the problem and upon the situation. Some observations can be made, however, which may be useful in helping to decide which is better in a specific situation. It may be well to make these observations by stating some questions and commenting upon them briefly.

Is a research staff located within the organization or is a research staff from outside the organization more likely to

- (1) be able to undertake and conduct the research with the greater objectivity?

Usually an outside research staff, particularly if it is from a well known research institution, is able to resist with greater success any pressures likely to result in a loss of objectivity.

- (2) be able to focus the research on the fundamental dimensions of the problems being considered?

Often the officers of an organization are so close to the problem involved that they tend to see only the immediate problems and often confuse symptoms and basic causal factors. The research is more likely to be productive if it is focused on the causal factors rather than symptoms. An outside research staff, particularly if it comes from a well known and respected research institution, is often in a better position to focus the research on the causal factors rather than symptoms or immediate operating problems of a transitory nature. An inside research staff of high prestige and power may, at times, be equally successful in being able to focus the research on the basic problems. But it is difficult for an inside research staff to question and do research on the assumptions and underlying philosophy upon which its top management is operating.

- (3) have a better knowledge of the problem and of all its ramifications?

Although an inside research staff may at times be handicapped by not being able to tackle a problem in terms of its more important dimensions, it nevertheless usually knows more than an outside research staff about the problems of an organization. An inside research staff

usually has an appreciable amount of information available about the problems of the organization, their history, and some of the major developments with regard to them

- (4) receive the full cooperation of all of the persons involved?
- An inside organization usually has closer contacts with the personnel whose cooperation will be important to the research. If an atmosphere of confidence and trust permeates the organization, an inside research staff is likely to obtain the better cooperation. On the other hand, if conflict and a substantial amount of fear and distrust are present, an outside organization with a reputation for objectivity and integrity is likely to be able to obtain the better cooperation. In general, any research operation which depends upon obtaining data from people's responses will obtain full, accurate data only when the people whose cooperation is required feel that they can trust the research staff.
- (5) have or be able to obtain the personnel required to do the research, including personnel with the specialized competence that may be required?

The desirability of creating within an organization a trained research staff including specialized technical personnel will depend in part upon whether this is a one time study or one of a continuous series of related studies. If the organization is continuously conducting a substantial amount of such research, it is often advantageous for it to create its own research staff. If it is a one time study, or if the volume of research undertaken is limited, the organization often obtains better and more economical research by using an outside research staff rather than creating one for the single study.

- (6) be able to exercise the influence required to have the results of the research effectively used?

In some situations an inside research staff has the power and prestige to assure that the results are likely to be used. In other instances an outside research staff with an excellent reputation is more likely to be able to exert the influence required. In deciding whether to use one or the other or both in a cooperative arrangement consid

eration can be given to the prestige and power required and the amount each is judged to have. Another dimension of the problem is probably at least as important as the question of an inside vs. an outside research staff. This question concerns the status of the person to whom the research staff reports. If he is a top officer, there is likely to be sufficient influence present to encourage serious consideration of the results. If he is a person of minor importance, the research findings are much less likely to receive consideration.

SELF-SURVEYS. The self-survey (38, 40) is a procedure which has merit where the problem is not too complex and where widespread participation is advantageous. With this procedure the organization itself conducts the study with appropriate technical advice and assistance. Usually a manual is available to tell how, on a step-by-step basis, the self-survey is to be conducted. Most organizations also obtain the assistance of a competent social scientist to serve as a consultant during the period of the survey. This consultant answers questions and provides advice on how to conduct the survey. His services include such tasks as helping to train personnel to do the tasks required, advising on analysis plans, and helping in the interpretation of the results. In addition to facilitating participation, the self-survey usually has the advantage of costing less than a study done by a research staff, or at least of involving less cash outlay. It also provides a means of conducting a study when the shortage of trained researchers would make it impossible to do any other kind of study.

The self-survey, however, can be used only on a limited variety of problems. They must be problems for which it is possible to develop a reasonably simple self-survey procedure which will, nevertheless, yield results of satisfactory accuracy. A continuous danger in using self-surveys is that the utilization of unskilled persons will yield results containing serious errors. In situations in which this is especially likely to occur or in which the consequences of such errors would be particularly disastrous, it would be unwise to use the self-survey.

HIERARCHICAL LEVEL OF THE RESEARCH STAFF. It has been observed that when a research staff is under the direct supervision of the person whose operation is being affected by the research, the

research is often discontinued. To function successfully, a research staff must report organizationally to the person who is superior to the man whose operation is affected by the research results. This principle seems equally applicable to both the physical and the social sciences, and for the same reason. As the results of research lead to improvements, the head of the operation most affected may feel threatened. Insofar as he feels threatened, he is apt to wish to have the research discontinued or the results ignored. When this occurs, he may discontinue it or bury the research findings if he has the authority to do so. Consequently, the research staff in an organization should neither report to nor be directly responsible to the person whose operation is directly affected by the research but should report to the next higher echelon.

There is another important advantage in having the research staff report to the next higher echelon. When the research staff reports to the man whose operation is affected by it, he may limit the scope of the problems studied or he may order it to produce findings to prove him right in what he is doing. Such restrictions make it impossible to conduct research effectively.

Although the research staff should report to a sufficiently high echelon in an organization to protect its stability and integrity, all of its activities need not be conducted through channels via this level. Usually it is well to use the formal channels in agreeing on the nature and scope of a research project, but after that is done there is a distinct advantage in the research staff's establishing direct and close contact and communication with all the echelons and groups involved in the research. Even in using the formal channels in agreeing on the nature of the research project, authority should not be used to force the research on the units involved. It usually is unwise to undertake research for or in a unit which does not genuinely want it. When people are forced to cooperate, they are likely to give distorted information.

Problems in Research Design

BASIC RATHER THAN SUPERFICIAL VARIABLES It is not at all common for an organization to request that research be done on some problem about which it currently is very much exercised and which it feels important. Often, however, when this problem is examined

it proves not to be the best problem upon which to do research. It may be a symptom rather than the underlying problem. It may be only one part of a fundamental problem. Or it may not be stated in dimensions which permit systematic, quantitative research.

The first task of a research staff is to diagnose the situation and to prepare a clear statement of objectives for the research. Discussing the problem fully with the officers and staff of the organization facilitates this diagnosis. Attempting to state the problem in dimensions based upon the best available theoretical conceptualizations also helps.

The final statement of the problem and of the objectives of the research must be acceptable to the organization. Often the discussions involved in diagnosing the problem lead to a recognition and acceptance of the problem as stated in the research objectives. Sometimes, however, this does not occur. If further discussions do not lead to an acceptance of the problem as stated, the research staff may have to start on a pilot or small-scale study devoted to a peripheral problem but one which the organization recognizes and about which it is much concerned. From the results obtained in this pilot study, the research staff usually can demonstrate the nature of the basic problem upon which the major research should be concentrated. Often the research staff itself gains a clearer understanding of the basic problem as a result of the pilot study.

RELATIONSHIP BETWEEN THEORETICAL AND APPLIED OBJECTIVES. It is impossible to emphasize sufficiently that research devoted to the operating problems of an agency need not and, if well done, will not be concerned with the symptoms of problems or with minutiae. Nor will it be concerned solely with finding specific answers to specific problems. Evidence is accumulating which points to the advantage of designing research dealing with a specific operating problem in such a way that the results can be generalized and applied to other related situations.

If the scientist doing research for an agency seeks only to find specific answers to specific problems, he is likely to run into serious difficulties. One difficulty, for example, is that there are so many specific problems that he will be hopelessly swamped. Another is that the cost of the research is likely to exceed the value of the specific answers. But particularly important is the fact that by the time the research has provided a specific answer to a particular

problem the situation will have so changed that the original problem is no longer the problem. New ones have replaced it (11). By inference this indicates that research designed to meet the long range problems of an organization will be more valuable and have greater application than fire fighting research designed to meet immediate problems.

The great value of generalized knowledge for dealing with specific problems was well stated in a remark attributed to Kurt Lewin. Nothing is so practical as a good theory. In the design of research focused on major variables the probability of significant findings is increased if the best available theory is used as a guide as to what to measure and what relationships to test. The better the theory used in guiding the research design the greater is the probability of finding marked and important relationships. Obviously the more that the research discovers about those major variables which have a marked relationship to the problem being studied the greater is the contribution of the research to solving this and related problems.

Generalizations or statements of principles which summarize the marked important relationships discovered in the research have two valuable uses. They serve as guides to help solve problems like the one upon which the research was focused. They also make a contribution to available scientific knowledge and the development of theory. Cartwright's *Some Principles of Mass Persuasion* is a good illustration (6-7).

In situations in which important changes in the character of the problem are likely to occur between the beginnings of the research and the availability of the results it is necessary for the researcher to take this into account in designing his research. In addition to concentrating research on the major variables involved he often will find it useful to design his research so that the results will adequately cover any reasonable range of change that may occur in the situation during the time required for the research. One way of doing this is to design the research so that it will yield results satisfactory for dealing analytically with two or more widely differing situations. If these assumptions involve situations more extreme than any that is likely to occur then the actual situation at the end of the study will fall between the extreme situations assumed. By bracketing his problem in this manner and having adequate data

to deal with a range of situations, the researcher usually is able to make valid and useful derivations from his findings to the situation that exists when the research findings become available

RESEARCH ON PRINCIPLES AND PROCESSES In organizational research (see Case A, pp 620-630), it is well for the research staff to emphasize and re-emphasize that the objective of the research is to discover the relative effectiveness of different methods and principles and that the study is in no way an attempt to perform a policing function. The emphasis must be on discovering what principles and methods work best and why, and not on finding and reporting which individuals are doing their jobs well or poorly.

Unless these objectives are made clear to all and rigorously adhered to, it will not be possible for the research staff to obtain the full cooperation that it needs from the people in the organization being studied. It is important for the research staff to make clear to every person that the interviews, questionnaires, and other data obtained from each person will be kept strictly confidential. People need to know that these materials are being collected for purposes of statistical analysis and that no one will be able to tell which specific answers were given by which individual.

The commitment of confidentiality must be clearly given and strictly adhered to. This at times may require the research staff not to report separately the data for very small groups, since to do so might reveal the attitude and answers of the individual members of the small group.

An orientation focused on discovering better principles and methods of organization and leadership reassures persons who may feel threatened by the research. If they feel that the research is to learn how to help them to do their job more successfully, they usually are eager to cooperate. This cooperation usually increases as they see the research results used for this purpose rather than to discharge or demote those whose work at present is not successful.

Assuring Use of Research Results

INDUCING COOPERATIVE RATHER THAN DEFENSIVE ATTITUDES Measurements of any commercial, industrial, or governmental operation almost always show that some things are being done well and other things are not being done so well. In examining these research

results, the officers of an organization can take primarily either a constructive or a defensive attitude toward the data. Fortunately, most officers take a constructive point of view. Occasionally, however, a company officer or supervisor takes a defensive attitude and immediately becomes fearful when data are obtained which show that his operation is not now functioning in the best possible manner. His impulse, as soon as he has seen such research results, is to lock them up immediately so that no one else can discover that the operation is functioning imperfectly. Most company officers or supervisors take the opposite point of view when looking at similar data. Their reaction is to look at the results which present a favorable picture with pleasure but to look at them hastily. They then turn with genuine enthusiasm to the results which indicate where and in what manner the operation can be improved. They immediately share this information with their colleagues, subordinates, and all other relevant personnel in order that the necessary steps can be taken which will lead to further improvement in the company's performance.

As we shall see, there is much that both the officers of an organization and the research staff can do to prepare and assist the personnel of the organization to take a constructive rather than a defensive attitude toward research results.

PARTICIPATION IN PLANNING AND IN INTERPRETATION If people are unfamiliar with a research project and know little about it, they are not likely to understand the findings or be interested in applying them. Personal involvement not only decreases the barriers to the use of data, it increases the probability that the results will be understood and accepted. Particularly important, however, it yields positive motivation to apply the results. This involvement should include all those who can influence the application of the results and should begin at the very outset of the project and increase as the project reaches the analysis stages. To wait until research results are available before attempting to obtain participation represents a failure really to use participation and is likely to lead to the full or partial rejection of the results.

The effectiveness of participation and involvement depends upon the rate or timing of the efforts devoted to this purpose. There seems to be no substitute for taking adequate time at many points in the process. The first point occurs when an organization

is considering whether to have research done on a problem it faces. If high pressure selling is applied, resistance is likely to occur. On the other hand, if the problem and needs of the organization are examined carefully and consideration is given to the help that research can and cannot provide, without pressure for a decision, the officers of an organization usually are more likely to understand and accept the assistance that research can probably provide. Also, the research staff is more likely to understand the problem and be more able to design an efficient study than if the decision to proceed with the study is made hastily. When an organization has decided after careful consideration that it will benefit from having research done on its problems and begins to press the research staff to have the research done, its officers are much more likely to be sufficiently interested in the study to take the time and energy required to become fully involved in the research.

Obtaining the participation of the relevant personnel in the planning stages of a study yields two dividends. It enriches and improves the material used in planning the study, and it also achieves the desired involvement. A similar gain occurs in using participation in the analysis and interpretation phases of the research project. The knowledge of company operations possessed by company officials and employees makes them experts whose help is needed by a research staff in planning a study and interpreting the data.

PRESENTATION OF PRELIMINARY FINDINGS AND ACCEPTANCE OF RESEARCH RESULTS The involvement and interest of the officers of an organization tend to wane if the research staff waits until the completion of the analysis before presenting any results to them. Moreover, people usually can do a better job of interpreting research results if they are given time to assimilate gradually the major findings emerging in the research. If nothing is reported to an organization until the final analysis is presented, the officers are confronted with a body of data which often includes some results which surprise them. The research staff has then faced them with what amounts to a 'take it or leave it' situation, and neither alternative is desirable from the standpoint of the researchers. On the other hand, when the research staff presents to the officers of an organization some preliminary inkling of the probable results but presents them as highly tentative, the officers are not compelled immediately to

implicit demand that the others make an immediate redefinition of the situation, serves only to increase the emotional resistance and the amount of time ultimately required to get the findings accepted and used. It is best to give the individual or group ways to save face—let them explore all of the different possible meanings which the findings might be assumed to have—before going ahead. One of the more important things that an outsider can effectively do here is to provide the individual with the motivation to re-examine his psychological field to see if there are not even better interpretations to the perceptual clues he has been getting and piecing together than the pattern which heretofore has satisfied him (32a).

(7) The results should be in simple, nontechnical language. Shop language and graphical presentation should be used as much as possible. This facilitates self-analysis by making the group realize that the data deal with their situation and are not something belonging to the research organization.

ANALYZING DATA IN GROUPS The procedure described in Case A (pp. 620-630) involves working with groups rather than with individuals. Lewin (19-20) and his students (10) have emphasized the power of the interacting forces exerted by group members on one another. Mann (31) finds that this is particularly applicable to a group's discussion of data dealing with its own work situation. Participation in group discussions and group decisions concerning future action sets into motion pressures for action which are more effective than when individuals alone are concerned. By working with and through groups, use is made of these continuing group forces.

The group situation seems to be important for several reasons.

(1) Through group discussions the findings can be examined in a broader perspective because the group brings to the data experience that is richer and more varied than that of any one individual.

(2) Group discussions, by allowing the pooling and exchange of this wider range of information, also provide the psychological situation in which superiors and subordinates at all levels can discuss possible solutions and thus give one another new and improved ways of viewing and solving their problems.

(3) The discussion by groups of the research data compels all members of the group to recognize openly the existence of the

major problems revealed by the data. Important and serious problems which may have long been festering are brought to light in an atmosphere which leads to constructive attempts to solve them.

(4) Group discussions also help supervisors at all levels to learn what is expected of them by the group concerning their relations with subordinates, associates, and their own chief.

(5) Group decisions concerning the next steps put powerful pressures in the form of reciprocal expectations on each member to carry out the decisions agreed to by the group.

USE OF HIERARCHICAL SOURCES OF INFLUENCE To assure effective application of research findings, it is important to recognize the hierarchical structure in an organization and it is also essential to utilize the power structures as perceived by members of the organization. Any series of meetings to achieve participation should follow a sequence which recognizes the presence of these forces. Research data presented to the groups in the meetings should show how different groups in the organization perceive the power roles of other persons in the line and staff groups. The people at the top of each organizational unit—particularly if they are perceived as competent and powerful—are found to exercise appreciably more influence on the organization than any other persons within it.

One reason why it is so important to secure the interest and full support of top management in planning research and interpreting the results arises from the peculiar problems involved in applying the results of social science research. The president of an automobile company does not have to understand the research involved in developing a new automobile engine when he approves plans for putting it into production. All he needs to know is that it will perform better and require less fuel than the present engine. The problem of applying social science research results, however, is not so simple. Effective application depends upon full understanding and use of the research by the top management of the organization. Consequently, it is important to keep top management fully informed of the progress of the research and fully involved in the application of the research findings.

Even in physical science and engineering research it helps to have top management interested in the research and identified with it. This interest encourages design engineers and those who must apply the results to follow the research more closely, and it reduces

accept or reject the data. Also, as further data are reported to them and progressively build a clearer and clearer picture of the results, the officers follow it with interest. During this period, they can test the validity of the results by using other evidence or clues. This testing and discovering that the results are valid facilitates their acceptance.

People find it difficult to make a major change in their thinking rapidly. It seems to require time for each of us to test new ideas and new results and gradually to discover their validity and to accept them. Not until then are we willing to build decisions upon these new findings. There seems to be no substitute for time in this process. Whenever pressure is exerted to achieve changes in points of view or in thinking in unduly short periods of time, there is likely to be strong emotional resistance to it.

SELF ANALYSIS TECHNIQUES AND THE USE OF RESULTS¹ Participation in the form of self analysis is more likely to be followed by changes than if the analysis is made by someone else. This point and the specific procedures used to implement it are not new. Clinical psychologists are well aware of the importance of self analysis for bringing about change (36). Some of the procedures which appear to be particularly effective follow:

(1) One important dimension of self analysis as used, for example in Case A (pp. 620-630) is that it starts from and centers around subjective measurements. This tends to keep the discussion in an objective atmosphere. There are no statements, reports, or recommendations by outside experts to which an individual or a group can take exception; there are only unemotional, objective data.

(2) In the series of meetings described in Case A, no expert tells a group what the data mean or what its problems are. The interpretations are worked out by the members of the group themselves. The representative from the research organization there to answer technical questions which the group leader cannot answer. Prior to the meeting, members of the research staff always study the data that are to be discussed. At times in the meetings the research staff members may ask questions about the data to help focus attention on what appear to them to be important points that

¹ The following three sections have been taken from Floyd C. Mann and Rensis Likert, *The Need for Research on Communicating Research Results* *Human Organization* 1952, 11, 4, 15-19.

are being overlooked. By taking this role, the research representative avoids making recommendations that may be naive in the light of the entire situation. He also side-steps many of the individual and group protective mechanisms which are set into action during any real evaluation of the self or the organization in which the self is deeply involved. The use of questions to help focus attention on the important problems and on the solutions suggested by the data produce solutions from the group and assure their acceptance by the group.

(3) Resistances have to be recognized and worked through, not glossed over. As might be expected, survey results are occasionally quite different from what the group expects. When this occurs, resistances in different forms and degrees of overtness have to be met and dealt with before the group can proceed again toward its objective. If the handling of resistances is postponed, they will surface later when it may be much more difficult to deal with them adequately or when it is too late to handle them at all. When resistances are not dealt with effectively, they are apt to block constructive action.

(4) Timing and pacing are important in facilitating the acceptance of the data and gaining recognition of the need to act upon them. In those situations in which the survey results are quite different from what is expected, it is necessary to proceed cautiously, preferably letting the individuals who are surprised set the tempo. This means letting the group pace itself in the speed with which it considers the different aspects of the findings and also in determining the depth to which the analysis and interpretation of the data will go at any one meeting. These two factors are important in that they tend to reduce the number of times that resistances arise because the group members are not yet prepared to understand or ready to accept certain findings as facts.

(5) It is well to present the results in a positive atmosphere, emphasizing first the results which show what is being done well. In presenting results dealing with operations which can be improved, it is well to orient the discussion toward what the data suggest are ways to make the operation better. Consideration of how to improve enlists interest, whereas concentration on weaknesses or failures produces anxiety and an avoidance reaction.

(6) Arbitrary insistence that the data are accurate, which is an

implicit demand that the others make an immediate redefinition of the situation, serves only to increase the emotional resistance and the amount of time ultimately required to get the findings accepted and used. It is best to give the individual or group ways to save face—let them explore all of the different possible meanings which the findings might be assumed to have—before going ahead. One of the more important things that an outsider can effectively do here is to provide the individual with the motivation to re-examine his psychological field to see if there are not even better interpretations to the perceptual clues he has been getting and piecing together than the pattern which heretofore has satisfied him (32a)

(7) The results should be in simple, nontechnical language. Shop language and graphical presentation should be used as much as possible. This facilitates self-analysis by making the group realize that the data deal with their situation and are not something belonging to the research organization.

ANALYZING DATA IN GROUPS The procedure described in Case A (pp. 620-630) involves working with groups rather than with individuals. Lewin (19, 20) and his students (10) have emphasized the power of the interacting forces exerted by group members on one another. Mann (31) finds that this is particularly applicable to a group's discussion of data dealing with its own work situation. Participation in group discussions and group decisions concerning future action sets into motion pressures for action which are more effective than when individuals alone are concerned. By working with and through groups, use is made of these continuing group forces.

The group situation seems to be important for several reasons:

(1) Through group discussions the findings can be examined in a broader perspective because the group brings to the data experience that is richer and more varied than that of any one individual.

(2) Group discussions, by allowing the pooling and exchange of this wider range of information, also provide the psychological situation in which superiors and subordinates at all levels can discuss possible solutions and thus give one another new and improved ways of viewing and solving their problems.

(3) The discussion, by groups, of the research data compels all members of the group to recognize openly the existence of the

major problems revealed by the data. Important and serious problems which may have long been festering are brought to light in an atmosphere which leads to constructive attempts to solve them.

(4) Group discussions also help supervisors at all levels to learn what is expected of them by the group concerning their relations with subordinates, associates, and their own chief.

(5) Group decisions concerning the next steps put powerful pressures in the form of reciprocal expectations on each member to carry out the decisions agreed to by the group.

USE OF HIERARCHICAL SOURCES OF INFLUENCE To assure effective application of research findings, it is important to recognize the hierarchical structure in an organization and it is also essential to utilize the power structures as perceived by members of the organization. Any series of meetings to achieve participation should follow a sequence which recognizes the presence of these forces. Research data presented to the groups in the meetings should show how different groups in the organization perceive the power roles of other persons in the line and staff groups. The people at the top of each organizational unit—particularly if they are perceived as competent and powerful—are found to exercise appreciably more influence on the organization than any other persons within it.

One reason why it is so important to secure the interest and full support of top management in planning research and interpreting the results arises from the peculiar problems involved in applying the results of social science research. The president of an automobile company does not have to understand the research involved in developing a new automobile engine when he approves plans for putting it into production. All he needs to know is that it will perform better and require less fuel than the present engine. The problem of applying social science research results, however, is not so simple. Effective application depends upon full understanding and use of the research by the top management of the organization. Consequently, it is important to keep top management fully informed of the progress of the research and fully involved in the application of the research findings.

Even in physical science and engineering research it helps to have top management interested in the research and identified with it. This interest encourages design engineers and those who must carry out the results to follow the research more closely and it reduces

the lag between the development of new knowledge and its actual use

USING DATA SO THAT IT PRESSES FOR ACTION Business executives and government officials who have made extensive use of social science research often point to the fact that they use the research results in such a way as to make the data do the work of pressing for action.

The Director of the War Finance Division during the last war used research results very effectively to press for action. One illustration will serve to show how he did this. He knew, for example, from his experience as Chairman of the X State War Bond Committee, that personal solicitation was essential if substantial sales of bonds were to be made to large numbers of individuals. He discovered, also after he became Director of the War Finance Division, that the chairmen of many other states did not accept *his* experience as a guide to what *they* ought to do. They shuddered at the thought of having to recruit and train tens of thousands of volunteers to serve as solicitors in war bond campaigns. Consequently, when he urged them to use solicitation and cited his own experience, they would solemnly assure him that their state was different from ——— state and that personal solicitation was not necessary to sell bonds in their state. The result was an impasse and, at first, several states did not use personal solicitation.

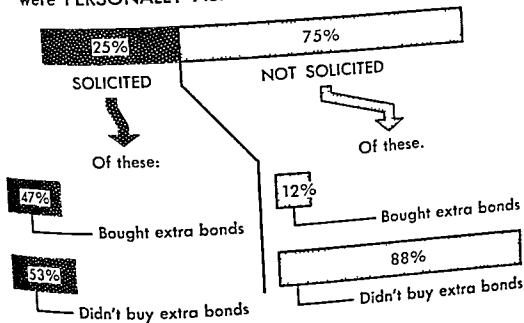
A study of the effectiveness of the Second War Bond Drive provided the director with data such as that shown in Figure 1. He had a brief pamphlet prepared which showed these and related results. He distributed these pamphlets to all the state and county war bond committees. He also had these results presented in the regional meetings in which the plans for the Third War Bond Drive were discussed and developed. The net effect of using the data in this way was that war bond committees convinced themselves of the value of solicitation. They recruited and trained a much larger group of solicitors. As a result, the number of people who were personally asked to buy war bonds increased from 25 percent in the Second Drive to 50 percent in the Third, and the sales of Series E Bonds almost doubled.

The impartiality of accurate measurements usually facilitates objective consideration of the facts, and this leads to the acceptance and implementation of effective policies. When decisions are made

by pitting one person's experience and judgment against that of another, there is usually disagreement. Often the best decision is not made, and any decision that is made is accepted and followed only half heartedly by many of the persons involved.

APPRAISING USE OF RESEARCH RESULTS BY REMEASUREMENT One characteristically human tendency is to assume that whenever a

Of all gainfully employed in the country, how many were PERSONALLY ASKED* to buy in second drive?



*Includes those whose wives were solicited

FIG. 1.

change is made it results in an improvement. Unfortunately, this is not always the case. There is only one way to know whether a change has resulted in an improvement, and that is to measure the effect of the change. It is not uncommon to find that the first application of research results has made no measurable improvement. It may take two or three attempts at change to produce improvement of any magnitude.

change be introduced generally, where applicable, so that all units will benefit and the organization derive maximum results from the change. A demonstration of the value of testing changes experimentally is often all that is required to convince an organization that usually it will derive greater benefits and derive them more rapidly if it tests the proposed changes experimentally before introducing them generally.

SOME ILLUSTRATIVE CASES

In considering how to organize and conduct research devoted to the problems of a specific organization, it is important to differentiate two broadly different kinds of situations. This is necessary since this difference affects the way in which the research should be organized and conducted if effective use of the results is to be obtained. One general type of situation is that in which the research deals with problems involved with the organization's internal operations. Case A below is of this type. In situations of this type the research operation is likely to have an appreciable impact on the organization and how it operates. Consequently, it is extremely important to have the full support and interest of the very top officers of the organization and to have their full participation in the plans for the research and its application. Unless top management understands, seeks, and fully supports the changes being undertaken, and understands the character of the resistance to the changes and the reasons for this resistance, the attempts to apply the research will encounter real difficulty and may even fail.

The second general type of research situation is that in which the research deals with problems faced by the organization which are outside of the organization. Studies concerned with the buying behavior of consumers or with citizen response to governmental activities illustrate this general kind of problem. Case B below is an example of this kind of research.

Case A (A Synthetic Case)

The Director of Industrial Relations in Company A approached a research organization one day with the statement that the com

pany had no immediate, pressing problem but that it had some long range problems which he felt would be benefited by research. After a lengthy discussion of some of these problems and a consideration of what kind of research design would be most likely to yield useful results, the company director and the head of the research organization agreed to give the matter further thought and have another meeting involving other persons from both organizations.

Further discussions of the potential value of research for helping to solve the long range problems, led the Director of Industrial Relations to recommend to the President that Company A proceed with the proposed research. This was followed by a meeting of the head of the research organization and two members of the research staff with the President, Executive Vice president and the Director of Industrial Relations of Company A. In this session, the potential values of the research project were discussed at length. Some discussion took place on how the research results might be applied and the problems likely to be encountered in this application. On the basis of this meeting and the recommendations previously made, the President authorized the study to be made. The research organization, before agreeing to proceed with the study, held a similar kind of meeting with the President and Executive Committee of the union local. The response and assurance of cooperation obtained in this meeting and the genuine interest in the research on the part of top management encouraged the research organization to agree to undertake the proposed study.

COMMENT Many organizations like Company A start research with long range objectives. Others start it initially as a result of more immediate problems. Excessive turnover or absenteeism, stealing and low productivity are a few examples of immediate problems which may lead a company to ask for assistance from a research staff. Often these more immediate problems are merely symptoms of fundamental problems and need to be so treated in the research design and in applying the research results.

The first step in undertaking the research was to organize a series of planning sessions. The purposes of these sessions were as follows:

- 1 To obtain as much information as possible from company officers and personnel as to the major dimensions of the long range problems. This information was needed to be sure that the research

design included all the major hypotheses that should be encompassed in the study

2 To inform all interested persons and groups about the study in a way that would enable them to ask questions about the study, to express any fears or reservations they might have, to discuss these fears and to obtain all the information they wanted or needed in order to have sufficient confidence in the study to cooperate fully with it

3 To plan and conduct the study in such a way that all persons whose action was required to implement the study were fully informed about it, were interested in it, and had an adequate opportunity to ask that the study obtain data they needed to make the best possible decisions on the problems falling within their areas of responsibility Taking sufficient time to secure this involvement was necessary for the results of the study to be implemented

COMMENT It is important to give all the people who are affected appreciably by a research project an opportunity to learn about it fully, to ask questions, and to make suggestions with regard to it Full access to information about the study and as much involvement and participation as is feasible reduces fears of the research and increases the likelihood that the results will be used constructively

The planning sessions were not compressed into too short a period of time and were arranged in such a way that those persons whose actions would be required to implement the study had an opportunity to participate and to raise the questions that were on their minds One of the important functions of these planning sessions was to stimulate thinking to consider possible courses of action which might improve present and future operations This helped to suggest objectives for the research by indicating the data required to help in a choice among possible alternative courses of action

Another important function which these planning sessions were designed to perform was to help managerial and supervisory personnel to become better aware of the character and magnitude of some of the larger problems, especially human problems, which they faced Often people are so immersed in day to day operations that they are not aware of some of their larger and more important problems

As the plans for the study proceeded and the actual collection of the data got under way, it was felt desirable to keep the company's personnel, whose involvement was important, informed of developments. This was done through special communications to them as well as through the reports in the company house organ. This sense of involvement was maintained and even increased by promptly reporting back some of the preliminary results of the research.

COMMENT Unless there are fairly frequent reports to interested persons and groups on the progress of the research, their interest tends to lag. Moreover, whenever anyone feels left out and uninformed, he tends to develop fearful and suspicious attitudes. Consequently, whenever a research staff fails to keep all those persons who have a relationship to a research project informed of developments and progress, suspicious attitudes toward the research are fostered. These fearful and suspicious attitudes, in the absence of accurate information, are apt to lead to inaccurate rumors about the research. The effect of these rumors upon the research can be devastating.

In presenting the preliminary results, emphasis was placed on their preliminary and tentative nature. Company officers were reminded that more detailed analysis might well change the interpretation of the initial results. An illustration that proved useful was to cite the experience of a company which found that those employees who were union members had a less favorable attitude toward the company than non union members. This relationship, however, was changed as soon as the character of work of the employees was held constant. Blue collar workers had a less favorable attitude toward the company than white collar workers, and when the effect of union membership was analyzed separately for blue collar workers and for white collar workers, the findings were changed. Union members then were found to have slightly more favorable attitudes toward the company than non union members.

Occasionally, the initial reaction of a company officer to measurements which showed that a particular portion of the company was performing appreciably less efficiently than the rest of the company was to feel that there should be a change in the supervisory personnel involved. To forestall this reaction, top management in Company A emphasized the responsibility of management to help

supervisors and managers to develop the skills called for in their jobs. There was continuous emphasis that the purpose of the research was not to perform a policing function but rather to find what principles and methods are associated with better performance and how to help mediocre units to improve. Throughout the study, the focus was on discovering the methods which produce successful operation and learning how to train people to use these methods effectively. It was important that supervisory personnel not fear the research as a policing operation likely to hurt them but view it rather as a major effort to help each of them to learn how to perform his job more successfully.

The results of the study in Company A fell into two distinct categories. The first kind of results dealt with the measurements of company operations, either for the company as a whole or for sub-units of the company. The other kind of data was based on cross tabulations which, for example, showed the different levels of employee productivity associated with different methods of supervision. This second kind of analysis, in addition to its value to Company A, adds to the body of fundamental knowledge about the principles of organizational functioning and leadership. Each of these two kinds of results was reported to the company personnel.

In Company A, the results initially presented to company officers were of the first kind. They were "straight runs" showing results for the company as a whole and for some of the major operating departments of the company. These and similar data were presented first to the President and the Executive Vice president in a meeting which was organized by the Director of Industrial Relations. The presentation of the results was done by the member of the research organization who was in charge of the study.

This presentation did not include a detailed report presenting an analysis and an interpretation of the data. On the contrary, only tabular results were presented and discussed. The discussion and interpretation were participated in by all of those present at the meeting. For example, questions were asked by the President as to why attitudes toward first line supervision differed so substantially in two departments. Tentative hypotheses were offered by him and the Executive Vice president as to possible causes. Tabulations already completed by the research staff proved some of these hypotheses untenable and corroborated others. Additional hypoth

eses that were suggested became guides for further analyses of the data and these further analyses were examined in subsequent meetings with these and other company officers

Further meetings with the President, Executive Vice president, and Director of Industrial Relations were held which were devoted primarily to looking at the results for the company as a whole. In the meetings, attention particularly was devoted to analyzing those results which suggested that there was need for some change in company policy. Several changes occurred, but only after further analysis of the data and discussions with those particularly affected. One change, for example, involved an important modification in some of the benefits provided by the company for employees. Another change resulted in a substantial shift in the manner in which job evaluation was performed. A third produced significant changes in the operation of the suggestion system.

At the end of the initial meeting, the President and Executive Vice president suggested that they would like to hold a series of similar meetings with groups of vice presidents present. Each group was to include departments having related activities.

At these meetings the data presented included not only the results for the company as a whole and for the departments whose vice presidents were present but also results for the major divisions within each of these departments. There were extensive examination and discussion of the findings by these groups of company officers. A major area of discussion in these meetings with the President and groups of vice presidents had to do with plans as to how each of the departments and their various subunits could best go about analyzing and applying the results of the study bearing upon their operation. It was decided to have each vice president present to his division heads the results for his department. In turn each division chief was to present to his section heads the results for the division and for the sections and so on down till the results were discussed with rank and file employees.

COMMENT In many instances these groups were considering problems of interpersonal and intergroup relations which had been points of friction for some time and which were emotionally loaded. Problems which had been avoided because they were extremely difficult were frequently brought out into the open by data. The objective impartiality of the findings helped the members of the

group to approach these problems in a constructive, problem solving orientation. Employees attitudes and feelings came to be facts instead of things to be disregarded because they appeared to be too difficult to handle or did not clamor for immediate attention.

One significant change was introduced when the results were reported by each vice president to his immediate subordinates. In the meetings with the President and the Executive Vice president, the results had been presented by the member of the research organization who directed the study. In the meetings conducted by each vice president for his staff, the responsibility for presenting the results was placed upon the vice president. It was necessary for him to become sufficiently familiar with the methodology used in the study so that he could describe the methods used and answer any questions about the study which might arise during the presentation of the results. For example, when some of the measurements indicated that a specific operation was not being done as well as might be expected, the people who were responsible for these operations often were likely to question the accuracy of the measurements. In these circumstances, it was necessary for the vice president to be sufficiently well informed about the methods used and the magnitude and scope of possible errors so that he could answer these questions as to the probable accuracy of the results. It was also necessary for him to point to other measurements and results which supported the specific result being questioned.

Each vice president was provided with tabulations showing the results for the company as a whole and for his department and for each division within his department. Results were available also for other related or comparable departments so that each vice president could show his group how they compared with other comparable units within the company. Data were also made available by the research organization showing measurements for comparable operations in other companies outside Company A. All these results made it possible for the vice president to discuss with his staff the pattern of results for his department.

In each of these meetings conducted by a vice president, a member of the staff of the Director of Industrial Relations was present to help answer any questions that might arise. In addition, a member of the research organization was also present to answer any specific questions about the research which the vice president might not be

prepared to answer, and also to explain what kinds of additional tabulations could be made. The groups were encouraged to ask for additional tabulations which would help them see what problems existed, where in their departments the problem occurred, what caused the problem, and what changes might be most likely to improve the situation. Most groups asked for these additional tabulations.

After these sessions with the vice president, each division head was asked to conduct similar meetings with his subordinates and they in turn asked their section heads to conduct meetings with their subordinates. In this manner the results were reported to lower and lower echelons in the company until the results had been reported to all the employees of the company.

COMMENT The series of meetings was started at the top of the line organization and worked down. It was found that in most departments in which the people at the top took a genuine interest in the findings, studied them, and tried to apply them, the data were discussed more adequately and used more constructively in working out action steps than where such interest was lacking.

A high degree of personal involvement in the analysis and interpretation was obtained through having each supervisor who was engaged in any managerial or supervisory activity participate in two kinds of meetings. First, there were one or more meetings in which he participated as a subordinate with his associates and under the leadership of his chief, and second, there were one or more meetings in which he participated as the chief of his group and conducted the meeting with his own immediate subordinates. This latter meeting compelled him to be familiar enough with the techniques used in the collection of the data and the over all results so that he could answer questions which arose in the discussion.

Some of the vice presidents asked their division heads and subordinates to report back to them the interpretations arrived at in the discussions and the actions that were taken on the basis of these interpretations. In those departments in which this was done, more effective use was made of the research results and greater action was taken on the basis of the results than in those departments in which the vice president asked his subordinates only to discuss the results with their staff and their subordinates.

The second major use of the measurements obtained in Com

pany A was to analyze the data to discover the pattern of relationships that existed between such variables as organizational structure and managerial and supervisory practices on the one hand and employee productivity and job satisfaction on the other

COMMENT Figures 2 and 3 show some of the relationship that were found and illustrate one of the methods of analysis used. Such findings as these are not only of value to "Company A" but they, and generalizations based on them and on similar results, make an important contribution to our fundamental knowledge about organizational functioning and leadership (15, 16, 17, 18, 24).

The pattern of relationships obtained from this analysis was used by Company A in three major ways

(1) These relationships were studied by the top officers of the company in staff meetings along with related data from similar studies. The purpose of this analysis by top officers of the company was to examine present company policies in the light of these relationships and to consider modifying those policies where the results suggested that a modification would bring improvement. All the different kinds of policies affecting employee behavior and reaction were studied, including policies dealing with such matters as wages, promotion, job evaluation, company benefits training and supervisory practices

(2) The training department built much of its material for training supervisors in human relations skills upon results obtained from the research. They used the research results for case material and for pointing to some of the more common human relations problems that existed in the company. In the training sessions supervisors discussed what changes in supervisory practices and procedures would be most likely to bring about improvement, especially in relation to those kinds of problems which, the data showed, occurred most frequently. They also discussed the patterns of relationships that were discovered to consider why certain kinds of supervisory behavior were associated with high or with low productivity or morale

(3) The patterns of relationship which were discovered, along with the related material from other studies, were made available to all levels of management as they examined and studied the specific measurements dealing with their own operation. The purpose of making these results available when the measurements were being

low-production section heads

are more closely supervised

than are high-production heads. . .

NUMBER OF FIRST-LINE SUPERVISORS	
Under Close Supervision	Under General Supervision

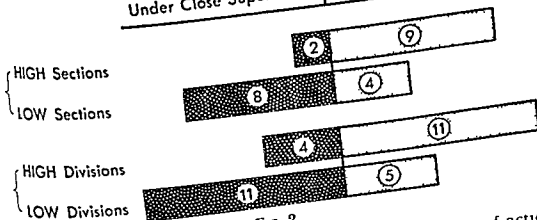


FIG. 2.

studied was to help guide the decisions as to which course of action among the alternatives present would be most likely to result in an improvement.

The officers of Company A encouraged middle and lower levels

"Employee-centered" supervisors

are higher producers

than "production-centered" supervisors. . .

NUMBER OF FIRST-LINE SUPERVISORS	
Production-centered	Employee-centered

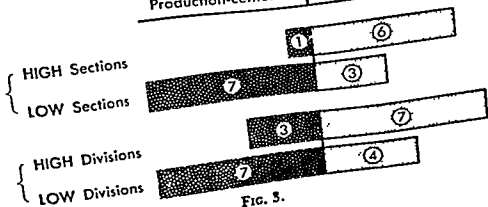


FIG. 3.

of management after careful study and discussion of the research results, to introduce modifications in their operation which the personnel involved felt would be likely to bring about an improvement. After these changes had been in operation for a sufficient period of time, top management authorized middle and lower levels of management to arrange with the research organization for whatever remeasurement was required in order to discover the effectiveness of the changes made.

This remeasurement, which was done in several different parts of the company, demonstrated that some of the changes introduced produced significant improvement and that other changes had a negligible effect. Some situations required two or three cycles of measurement, analysis, attempted change and remeasurement before improvement began to occur.

Case B (Research on a Population Being Served by an Agency)

Case B represents another type of situation. It differs from Case A in that the research is focused on persons outside the organization. In both cases the results were used to bring about improvement in the operation of the organization. In Case A, the changes involved internal relationships, in Case B the changes were program changes to help the organization achieve its objectives more effectively. In both types of cases, those persons who are in positions of influence need to participate in planning the research and in decisions on applying the research. Unless this is the case the research results are not likely to be applied.

In the summer and fall of 1941, the U. S. Department of Agriculture asked the Agricultural War Boards in the Great Lakes dairy states to undertake campaigns to increase the production of milk. This increase was necessary to meet the increasing demand in the United States as a result of increased purchasing power on the part of consumers and also to meet our commitments to England to supply substantial amounts of evaporated milk, cheese, and dried skim milk.

The officers in the Department of Agriculture who were responsible for this increase in milk production were eager to give the Agricultural War Boards in the Great Lakes states every assistance possible in order that their campaign to increase milk production

would meet with success. As a part of the assistance being given to the Agricultural War Boards, they asked the Division of Program Surveys in the Bureau of Agricultural Economics whether it would be willing to help the X State Agricultural War Board by conducting a study to help guide its campaign to increase milk production. After the Division had indicated its willingness to make the study, the Department of Agriculture officer responsible for the milk program asked the Chairman of the Agricultural War Board in X State whether he would like this assistance. The Chairman was interested in obtaining all the help he could on the important and difficult task he faced and was glad to have the study undertaken.

COMMENT This situation illustrates a difficult problem often encountered in the conducting of social science research for an organization or agency. In this case an arm of the federal government asked the research staff to make a study to help an agency of a state government. Even though this state agency had close functional ties with the federal agency involved, it was nevertheless proud of its autonomy and determined to maintain it.

In such circumstances, if the research staff had appeared to be responsible solely to the federal agency and interested only in the problem as seen by the federal agency, animosities and frictions would probably have developed which could have created a situation making the research difficult to conduct or causing the results to be ignored. To achieve the cooperation, support, and involvement needed, it was necessary for the research staff to approach the local agency with a genuine and sincere interest in learning what the officers of the agency felt their problems to be and what they would like to see studied.

This same kind of problem, of course, is faced whenever a central staff of an organization suggests that research be done for one of the operating units. It even occurs in a less obvious form when a company asks to have research done for a particular department. In all these situations the research staff must establish a cooperative relationship with the local people concerned which will develop confidence in the researchers and involvement in the research.

When members of the staff of the Division of Program Surveys met with the X State Agricultural War Board to discuss the proposed study and to develop specific objectives for the study, it became clear that the War Board wanted to get all the help it could but it was

not sure that research could help it or in what way. The members felt that, by and large, they knew the situation in X State well and that their campaign plans were sound.

The X State Agricultural War Board believed the following to be the facts:

(1) The State Agricultural Statistician's data for X State showed that the total number of cows in X State was larger than at any previous time in the state's history. Consequently, the War Board was of the opinion that virtually all the barns in X State were full and hence no further increase in number of cows was practicable at this time without an increase in barn space. The members also knew that their own barns were full and that the barns of all the farmers they knew were full.

(2) They knew that the price of milk in relation to the price of feed made it highly profitable to feed milk cows heavily, including grains and protein concentrates, in order to increase the milk production per cow. They were following this practice themselves, and all the farmers they knew were following this practice. They assumed that this was true among farmers generally in X State.

Believing these to be the facts for X State, the War Board was of the opinion that the way to increase dairy production was (1) to assure an adequate supply of feed grains and protein concentrates at a reasonable price, (2) to facilitate the building of additional barn space, and (3) to increase the available farm labor for dairy operation. Consequently, the War Board proposed to ask that feed be made available at a specified price and that priorities be granted or allocations made so that farmers could obtain all the lumber, concrete, steel, plumbing, and other material required for building additional barn space. The members also felt it desirable that steps be taken to increase the farm labor supply in X State.

With this background, the study was undertaken. The major objectives were to find the extent to which farmers were producing the maximum amount of milk and the steps which could be taken to make possible a further increase in dairy production. These objectives included discovering what resources, if any, farmers felt they needed to increase their dairy production. Such resources were additional barn space, milking machines, other equipment, more feed, higher quality feed, more labor, etc. The study also had as an objective discovering the extent to which dairy farmers in X State were

motivated to attempt to produce a maximum amount of milk and what the influences were that were motivating them to increase dairy production and what the motivational forces were that were acting in the opposite direction. This included seeking to discover how farmers felt about milk prices and the extent to which farmers knew that a price guarantee had been made on milk and dairy products through June 30, 1943. It was also desired to know the extent to which farmers knew that there was an urgent demand for an increase in milk production and the reasons for this increase in demand.

The study was designed with a sufficiently large sample so that the results could be analyzed separately for each of three major milk producing areas. These areas were the counties producing milk primarily for (1) cheese, (2) evaporated milk, and (3) dried skim milk. In each area a cross section of farmers and a sample of township AAA committeemen were interviewed. Interviews were conducted in a total of nine counties, three counties in each of the three areas. In each of these counties the three county AAA committeemen were also interviewed.

During the time of the interviewing the study director made an effort to drop in periodically for short visits with the Chairman of the State Agricultural War Board. He used these visits to report on the progress in the interviewing and to quote some of the answers being obtained in the interviews. He knew that many of the answers were proving to be quite different from what the War Board expected, and he wanted to prepare them for the results.

The results obtained from the interviews with the farmers and the AAA committeemen proved to be quite different from what the State Agricultural War Board expected. Their expectations were based on the statistics and reports available to them and what they learned from talking with their friends and acquaintances. But unfortunately, they lacked important information. They did not have available any data reflecting the information, attitudes, and behavior of the rank and file of farmers in a State.

The results, based on a cross section of farmers to quote the original report, were

1. There is ample barn space to accommodate considerable expansion in herds. Thus to 93 percent of the farmers.

tacted, the lack of barn space is not a determining factor in deciding whether or not to expand . . .

2 Labor supply in the state appears to be adequate, and the fear of a labor shortage does not seem to be an important obstacle to an expansion in production . . .

3 Equipment needs appear to be no handicap to increased production. . . .

4 An overwhelming proportion of X State farmers are satisfied with the present price of milk . . .

5 Running through all the data is the very noticeable thread that the farmer is uncertain, that he lacks confidence in the continuation of good prices, that he is apprehensive of a collapse of prices and markets after the war

It appears that memories of the catastrophe that struck farmers after the last war and the experience of the recent hard, lean depression years have sensitized farmers to the danger of debt

The great importance of guarantees of security and price over a long period of time is apparent . . .

6 A third major kind of brake on expansion is a distinct lack of information First, farmers are confused and only partially informed about the actions of the government in regard to production and prices Second, they lack knowledge about the possibility of increased output through better feeding practices . . .

7 Expansion of production in dairying can, of course, come in one or both of two ways larger herds and better feeding The first involves physical expansion, which may be hampered by the fears and uncertainties indicated above The latter, better feeding, could be done despite uncertainties However, better feeding is not the means by which the average farmer will make large increases .

Ninety four percent of those expanding production are enlarging herds The overwhelming proportion of those interviewed felt that they were already at, or close to, the economic limit on feeding intensity.

The results from interviews with township AAA committeemen showed a significantly different pattern from that obtained from

farmers The following results quoted from the original report present some of the major differences that were found

1 Committeemen seem to be responding much faster than other farmers to the present favorable conditions by planning to feed their cows more heavily this year

2 There appears to be some evidence that committeemen anticipated the need for increased production of milk and are taking advantage of it by increasing their herds fully a year in advance of other farmers

3 Strongly affecting the foregoing is the difference in information between the two groups Three quarters of the committeemen signified that they had heard of the government guarantees as opposed to only one fifth of the farmers

The interviews in this study were made between September 20 and October 1, 1941, and preliminary results were presented to the Agricultural War Board on October 2 The War Board was quite surprised by the results because the findings were almost the opposite of what it had believed to be the situation After it had the results available, it began to check them by talking to county AAA committeemen and various other groups of farmers After a few days of these discussions, the board became increasingly convinced that the results presented to it in the preliminary findings were substantially correct

The pattern that the board found was the one indicated by the study the well informed farmers, like themselves, knew of the increase in demand for milk, the reasons for this increase, and the probable stability of the increased demand These farmers, like the members of the State War Board, had already increased cow numbers to the point where their barns were full, and they were now increasing production further by heavier feeding of grains and concentrates County AAA committeemen displayed much the same pattern of behavior as the members of the Agricultural War Board Township committeemen were somewhat more like the rank and file of farmers than the county AAA committeemen but they too were ahead of the rank and file of farmers in the amount of information they had and in the extent to which they were already increasing milk production

COMMENT The research staff deliberately made much use of the pattern of results that was found. It pointed out that the township committeemen's attitudes and behavior were more like the results which the War Board had expected to find than were the attitudes and behavior of farmers generally. This helped to show that informed farmers were behaving as the War Board believed all farmers were behaving. The War Board members also recognized that their own friends and contacts tended, like themselves, to be much better informed than farmers generally. This made it easy to suggest that probably, as the results indicated, many farmers were not so well informed as the War Board had previously assumed.

After several days of studying the findings and testing them by talking with farm leaders and groups of farmers, the Agricultural War Board became convinced of the accuracy of the findings. It then drastically revised its plans for the campaign to increase milk production in X State. The revised plans called, first, for a major informational campaign through the mass media, especially radio, in which there would be much emphasis upon

(1) the great increase in demand for milk and milk products the reasons for this increase and the reasons why this increase in demand would not collapse suddenly

(2) the price support and guarantees to which the government was committed through June 30, 1943

(3) the importance of increasing dairy production primarily by the heavier feeding of grains and concentrates

In addition to this intensive campaign through the mass media the plans also called for meetings of farmers by school districts in local schoolhouses. These were to be evening meetings on a date which was to be about the first of November. At these school district meetings, the township committeemen were to explain and discuss with the farmers the need for the increase in milk, the character of the government price supports and guarantees and the best methods for achieving this increase in production through heavier feeding.

A final part of the entire campaign was to have each farmer called upon by a township AAA committeeman. The purpose of this visit was further to encourage farmers to increase dairy production and to answer their questions about the need for this increase. The

visit was to be combined with the usual Agricultural Conservation Program visit in which farmers were asked annually to sign up for approved soil conservation practices. This visit for the combined purposes was to be held earlier in the crop year than the annual conservation practice visit.

It was recognized by the State War Board that the township committeemen would have to be provided with full facts about the need for increased milk, the government program price guarantees with regard to milk, and the advantages of increasing production by heavier feeding. Material covering these points was prepared by the War Board and supplied to all county agricultural officials. Plans were also drawn for the visits that the township committeemen were to make to each farmer in order to explain the need for increased dairy production.

In order to test the adequacy of this plan, the materials prepared, and the training given township committeemen, three experimental counties were selected. During the period of October 14 to October 18 interviewers of the Division of Program Surveys accompanied township committeemen while they made their calls on farmers in these counties.

Several deficiencies were observed in the manner in which the township committeemen conducted the visits with farmers. One deficiency was a lack of information about the entire program for increasing milk production. Township committeemen, because they lacked this information or familiarity with it, tended to talk more about the soil conservation program, which they knew well, and much less about the milk campaign, which was the urgent requirement at the time. The specific deficiencies in information and in the manner in which the visits were conducted were reported by the Division of Program Surveys to the Agricultural War Board. The War Board instituted local training programs to overcome these deficiencies before the township committeemen called upon farmers during the early part of November throughout the milk producing areas of X State to encourage them to increase milk production.

COMMENT This is a very limited illustration of how the research action research cycle can be operated. After the initial study, plans for applying the results were developed and the adequacy of these plans was then pretested in an experiment or pilot project. The results obtained were then appraised through measure

ment, and the plan was modified and improved before more general application. When the problem is a continuing one, further measurements should be obtained after the general application.

The X State War Board found the assistance given to it in this study of so much value that the Agricultural War Boards of the other Great Lakes dairy states, when they learned of the study, asked for similar assistance in their own states. Time and financial limitations made it impossible to conduct separate studies for the War Boards in each of the states but representatives of the Division of Program Surveys met with these War Boards and described the results of the X State study, the major conclusions which came from it, and the specific applications that were made. Each of these State Agricultural War Boards then made corresponding applications in its own state.

An indication of the value of this project can be obtained by examining the increase in dairy production that occurred in X State during the ensuing twelve months. The increase in milk production for the year November 1, 1941 to October 31, 1942 over the preceding twelve month period was 6.7 percent. This was one of the highest rates of increase that occurred throughout the entire prewar and war period. Moreover, the study led the State War Board not to request the allocation of steel, lumber, cement, plumbing material, etc., which had originally been contemplated. This conserved scarce material for urgent needs in the war effort.

Case C (A Story of What Not to Do)

Case C, a condensed synthetic case, focuses on undesirable practices; reference to desirable practices has been largely omitted. It is presented here because there continues to be evidence that, among business, government, and other agencies, attempts are made to use social science research in situations in which the researchers are isolated from contact with operating personnel. Such conditions prevent participation and involvement by operating personnel in the research process and prevent the research staff from developing a full understanding of operating problems.

This organization was essentially a service research unit for operating governmental agencies. It was headed by an administrator who was competent but who happened to have no scientific training.

or research experience. This deficiency made it impossible for him to judge the competence of the research personnel on his staff and to know to whom and to what extent he could delegate tasks and functions. It handicapped him seriously also in his attempts to judge whether a proposed research project or program was the best possible design or would even adequately meet the requirements of the operating agency requesting the research.

This administrator was an earnest man who took his responsibility seriously and recognized that research results might exercise a major influence in decisions on important matters. He also recognized that some of the methods of the social scientists were newly developed. This caused him to feel that they were untested and hence likely, at times, to yield erroneous results. As a consequence he established procedures involving checks and balances.

Unfortunately, in order for his system to work, isolation of the research personnel from the operating personnel was required. This imposed a condition on the research organization which made it virtually impossible for it to perform the research services expected of it.

An agency desiring to have research done on a problem it currently faced could request to have the research done but only through a liaison person designated to perform this function by the administrative head of the research organization. A research unit or section was then given the research request through channels. The research staff then designed a research project to meet the problem as it was described to them through these channels. Usually there was little opportunity to discuss the problem or the proposed research with the agency requesting the study.

The absence of any except seriously restricted contact between the research personnel and the personnel of the operating agency tended to result in the research organization's either being asked only to make spot checks or to do research on relatively inconsequential problems, or not to be asked to do research. This occurred because the liaison person, not understanding the potentiality of the research methodologies, failed to see and to suggest to the agency personnel how some of their major problems could be tackled through research.

COMMENT *The personnel of operating agencies usually have no way of knowing the extent to which social science research can contribute to the solution of the many problems they face. They*

obtain this information primarily by discussing their problems sufficiently with social scientists for the scientists to become familiar with the problems and to indicate what kinds of assistance research might be expected to contribute. Consequently, in situations such as this, in which the contacts between the operating personnel and the research staff are restricted to a minimum, there is little opportunity for the scientists to become familiar with operating problems and to suggest how research might be used to help solve them. When the personnel of operating agencies are unable to obtain ideas and suggestions from the scientists, they necessarily limit their requests for research only to the very few possibilities which their restricted knowledge of research and its potentiality suggests.

Under such conditions, it is virtually impossible for the operating personnel to grasp the full magnitude of what research can do for them and to seek this assistance. Similarly, the research personnel without knowledge of the operating problems, do not recognize the need that operating personnel have for research and so are unable to suggest possible research projects. In these circumstances, research is done only on minor aspects of problems or on superficial problems.

After a request for research had been given the research staff, a study design was developed. The study design was then reviewed by one or more members of a group of research consultants which this administrator had appointed to advise him. These consultants usually spent only one or two days a week with the organization and therefore had only a cursory knowledge of the problems of the agency and consequently of the required research. Their discussions and advice usually resulted in some modifications in the research design. After these changes the consultants generally gave their approval which in turn resulted in the administrator's authorizing the staff to proceed with the research project.

COMMENT *Whenever research plans are reviewed by outside experts who do not have sufficient time to become thoroughly familiar with the problem, the research staff generally feels under pressure to use a research design which will be subject to a minimum of criticism and change. This, of course, results in a tendency to use a design of a traditional character rather than one which is specifically and efficiently focused on the problem.*

Upon completion of the research, a technical report was prepared by the research staff. This report, again, was reviewed and criticized by the research consultants and was then revised to take these criticisms and suggestions into account. A simplified, non-technical report was then prepared for submission to the agency which requested the research.

When this nontechnical report was ready, it was taken by the liaison person to the agency and discussed in a general way with the key operating personnel in the agency. The liaison person had not had research training and had only limited research experience. He usually had had no direct contact with the research which was done for this agency. As a rule, the only contact the liaison person had with the research was to participate in the initial discussion with the agency personnel in which the research was requested and in the editorial discussions which occurred when the simplified, non-technical report was being reviewed and put into final form.

The administrator in charge of the research organization did not wish the research staff to participate in the application discussions, any more than he wished them to participate in the initial discussions in which the research was requested by the agency. As a result, participation was rare.

The liaison person usually did not fully understand the research, was unfamiliar with the methods used, and had no appreciation of the limits of error within which the results could be interpreted or generalized. Often this liaison person had only a general impression as to what the research results meant and what they did not mean. He was of little help to the agency personnel when they would ask, "In relation to this problem, we can pursue policy *A* or policy *B*. Are we right in interpreting these results to mean that we should pursue policy *B* for these and these reasons?" The liaison person usually was unable to answer such questions because he did not understand the research well enough.

Since the liaison person was unfamiliar with the methodology used and the details of the research, he could not perform another important function involved in helping an agency to use the results of research conducted on its problems. Often as the officers of an agency examined the results of a study for implications for operating decisions, questions were raised which were not fully answered by

obtain this information primarily by discussing their problems sufficiently with social scientists for the scientists to become familiar with the problems and to indicate what kinds of assistance research might be expected to contribute. Consequently, in situations such as this, in which the contacts between the operating personnel and the research staff are restricted to a minimum, there is little opportunity for the scientists to become familiar with operating problems and to suggest how research might be used to help solve them. When the personnel of operating agencies are unable to obtain ideas and suggestions from the scientists, they necessarily limit their requests for research only to the very few possibilities which their restricted knowledge of research and its potentiality suggests.

Under such conditions, it is virtually impossible for the operating personnel to grasp the full magnitude of what research can do for them and to seek this assistance. Similarly, the research personnel, without knowledge of the operating problems, do not recognize the need that operating personnel have for research and so are unable to suggest possible research projects. In these circumstances, research is done only on minor aspects of problems or on superficial problems.

After a request for research had been given the research staff, a study design was developed. The study design was then reviewed by one or more members of a group of research consultants which this administrator had appointed to advise him. These consultants usually spent only one or two days a week with the organization and therefore had only a cursory knowledge of the problems of the agency and consequently of the required research. Their discussions and advice usually resulted in some modifications in the research design. After these changes, the consultants generally gave their approval which in turn resulted in the administrator's authorizing the staff to proceed with the research project.

COMMENT Whenever research plans are reviewed by outside experts who do not have sufficient time to become thoroughly familiar with the problem, the research staff generally feels under pressure to use a research design which will be subject to a minimum of criticism and change. This, of course, results in a tendency to use a design of a traditional character rather than one which is specifically and efficiently focused on the problem.

COMMENT This case suggests that social scientists will be well advised to counsel administrators against establishing an organization like that in Case C. They would be wise also to avoid positions on any research staff which is organized like that in Case C. Such an organizational arrangement imposes such restrictions on the research that it is virtually impossible to do significant studies and to have any research results used effectively.

RESEARCH AND TRAINING IN APPLYING RESEARCH

One of the important but neglected research areas in the social sciences is research on the process of applying research findings. The content of this chapter has been based largely on crude, unsystematic observations and experience, and it is therefore necessarily tentative. Systematic research is needed urgently to provide an integrated body of knowledge on how best to apply knowledge, especially on how to apply the knowledge emanating from social science research. There cannot be an applied social science technology of any magnitude and importance until a more systematic body of knowledge is available on how to apply research findings.

A review of current programs of graduate training for social scientists suggests also that there is need to broaden the curriculum beyond the current teaching of research methodology to include more systematic training in some of the methodological skills useful in the application of science.

The potential contribution of the social sciences to the solution of social problems is tremendous and largely unrealized. It is important not to oversell the potential value of social science research, either generally or for specific situations. It is equally important, however, for social scientists neither to undersell social science research nor to lack confidence in the scientific method as a tool for solving the complex social problems which our world faces. This faith underlies all the successful applications and uses of social science research.

the original analysis. Frequently, however, the original data could be retabulated or analyzed further so as to yield significant facts to help answer the new questions. But the liaison person was unable to suggest that this could be done because he was unfamiliar with the methodology and the data.

COMMENT: If research for an operating agency is done well, it will yield generalizations applicable to many problems and policies and not just specific answers to specific questions. But if these generalizations are to be used fully and correctly, it is essential that they be understood by the operating personnel. An effective understanding of the generalizations usually is best acquired by discussion of their implications for specific operating problems. This requires discussions of such problems by key operating persons with scientists who either have conducted the research or know it well. When administrative regulations prevent such discussions, the use of the research results is seriously limited.

There is another kind of problem which occurs when a liaison person without research training is placed between the research and the operating personnel. Often the most intelligent application of the research findings involves recognizing the presence of patterns of results and not merely considering individual items. Moreover these pattern interpretations may require knowledge not only of the findings from this current study but knowledge of the results of other research, including work done previously in this agency and work done elsewhere and published. The scientists who conduct the research have this background knowledge of other research results. A liaison person who lacks scientific training almost never has this knowledge and consequently is less likely to be able to help the operating person to see the patterns of results that have relevance for their decisions.

As might be expected, operating agencies found that research conducted in these circumstances was not very useful to them. Progressively they asked for and used research less and less.

This research organization was finally abolished. Some parts of it, however, were shifted to other agencies. Some of these, after they were shifted, managed to develop a close, functional relationship with the operating agencies whose problems they were studying. These parts survived and performed important and valuable research functions.

BIBLIOGRAPHY

- 1 Benne K D and Swanson G E The problem of values and the social scientist *J Soc Issues* 1950 6, No 4 27
- 2 Blakeslee A L Psychology and the newspaper man *Amer Psychologist* 1952 7 No 3 91 94
- 3 Campbell A Attitude research in the Department of Agriculture In Blankenship A B (ed) *How to conduct consumer and opinion research* New York Harper 1946
- 4 ——— Measuring public attitudes *J Soc Issues* 1946 2, No 2 58 66
- 5 Cartwright D Basic and applied social psychology *Phil Sci*, 1949 16 No 3 198 208
- 6 ——— Some principles of mass persuasion *Hum Relat*, 1949 2, 253 268
- 7 ——— Achieving change in people some applications of group dynamics theory *Hum Relat* 1951 4, No 4 381 392
- 8 Chein I Cook S W and Harding J The use of research in social therapy *Hum Relat* 1948 1, 497 511
- 9 Coch L and French J R P Jr Overcoming resistance to change *Hum Relat* 1948 1, 512 532
- 10 Festinger L *et al* *Theory and experiment in social communication* Ann Arbor Research Center for Group Dynamics 1950
- 11 Finch G Organization and opportunities in service programs of psychological research *Amer Psychologist* 1952 7, No 5 153 157
- 12 Hauser P M Social science and social engineering *Phil Sci* 1949 16 No 3 209 218
- 13 Jacobson E Kahn R Mann F and Morse N (eds) *Human relations research in large organizations* *J Soc Issues* 1951 7, No 3 1 71
- 14 Jaques E (ed) *Social diagnosis and social therapy* *J Soc Issues* 1947 3 No 2
- 15 Katz D and Kahn R Human organization and worker motivation In Tripp L R (ed) *Industrial productivity* Madison Industrial Relations Research Association 1951 pp 146 171

complex organizations a series of studies in a public utility Paper to be published in *Ohio Valley Sociologist*, 1953

- 31 Mann F C, and Baumgartel, H *Survey feedback experiment an evaluation of a program for the utilization of survey findings* Monograph in preparation
- 32 ———, and Likert R The need for research on communicating research results *Human Organization*, 1952, 11, No 4, 15 19
- 32a Mann F C, and Lippitt R (eds) Social relations skills in field research *J of Social Issues*, 1952 8, 3
- 33 Marrow A J *Living without hate* New York Harper, 1951
- 34 Merton R K The role of applied social science in the formation of policy a research memorandum *Phil Sci*, 1949, 16, No 3 161 181
- 35 Murphy G (ed) *Human nature and enduring peace* New York Houghton Mifflin 1945
- 36 Rogers C *Counseling and psychotherapy* Boston Houghton Mifflin, 1942
- 37 Shils E A Social science and social policy *Phil Sci*, 1949 16, No 3 219 242
- 38 U S Department of Agriculture *A manual for community self surveys of knowledge about nutrition* Washington D C Division of Program Surveys 1944
- 39 Watson G (ed) *Civilian morale* New York Houghton Mifflin 1942
- 40 Wormser M H and Sellitz C *How to conduct a community self survey of civil rights* New York Association Press 1951
- 41 Zander A Resistance to change—its analysis and prevention *Advanced Mgmt*, 1950 15, No 1, 9 11

INDEX

- Abel, T., 302, 304, 323
 accounts of small-group process, 305
 action research, 99, 637
 Adkins, D. C., 252, 296
 Adorno, T. W., 78, 96, 270, 272, 290, 296, 330, 379
 "after-only" experiment, 115
 agreement, observer, 391, 399-401, 410-413
 Allport, F. H., 148, 170
 Allport, G. W., 248, 297, 301, 303, 304, 430, 426, 467
 analysis
 group, of data, 616-617
 latent structure, 277
 methods of, 513-532
 in organizational research, 627
 of qualitative data, 421-467
 of results of field studies, 90-94
 strong method of, 471, 486-487
 survey, 33-36
 of variance in field studies, 91
 (see also content analysis; data, analysis, of; internal consistency; item analysis)
 Angell, R. C., 24, 52, 301, 304, 305, 317, 323
 Annis, A. D., 102, 103, 133
 anthropological approach, 59-62, 64-65
 application, of research, 590-602, 611-621, 628-630, 636-638
 applied research, 38-39, 99-100, 107-108, 169, 608-610
 area sampling, 186-187, 230-235
 Arnheim, R., 433, 467
 Arsenian, J. M., 413, 417
 Asch, S. E., 142, 170
 assumptions
 a priori, 536
 in factor analysis, 277-278
 normality, 530, 537
 of ratio scales, 280
 in sample selection, 189
 in scaling, 506
 attitude questionnaire, score on, 530-531
 attitudes
 census data and, 314
 as content of surveys, 32-33, 329-330
 content analysis and, 432-433
 expectations and, 150
 experimental, 582
 items in testing of, 512, 523-524
 toward organizational data, 611-612
 public and private, 62-63, 72-73
 scale of, construction of, 520-521
 study of, by interview, 329-330
 study of change of, 122
 unreliability in reporting, 42-43
 attraction, to a group, 150-151, 158-159
 attributes, latent, 494, 513-519, 528
 authoritarian personality, 78, 272, 290
 autobiographies, 304
 autocratic atmospheres, 138, 154, 156
 Back, K., 65, 96, 102, 133, 143-144, 150, 156, 158-159, 166, 170, 336, 379
 Baldwin, A. L., 430, 446, 476
 Bales, R. F., 389, 391, 411, 412, 414, 417, 440, 467
 Bancroft, G., 343, 379
 Barber, K., 140, 171, 433, 468
 Barnard, C. I., 57, 95
 Barron, Margaret, 377, 379
 Barton, A. H., 438, 442, 457, 469
 Baumgartel, H., 616, 646
 Bavelas, A., 116, 166, 170, 376, 379, 407, 417, 645
 Bechtoldt, H. P., 287, 297
 before and-after design, 24-25
 Belknap, G., 38, 52
 Benne, K. D., 119, 644
 Bennett, J. F., 277, 297, 505, 518, 533
 Berelson, B., 28, 54, 424, 425, 426, 428, 433, 437, 445, 446, 453, 456, 467
 Berkowitz, L., 383, 391, 401, 403, 417
 Bernaut, E., 429, 469
 Berson, M. P., 563, 576
 bias, 331, 339, 427
 (see also error)
 biograms, 304-305, 308
 Blakeslee, A. L., 644
 Blankenship, A. B., 342, 379
 Block, Jack, 100, 133
 Block, Jeanne, 100, 133
 Blumenstock, D., 434, 463
 Blumer, H., 301, 302, 323
 Bogue, D., 312, 326
 Bradford, L., 102, 163, 171
 Brat, D. W., 162, 171
 Bridgman, P. W., 476, 533
 Britt, S. H., 429, 467

Bronfenbrenner, U, 504, 533
Bruner, J S, 305, 323, 426, 467
business activity data, 310
business conditions, indices of, 316

Cahalan, D, 46, 52
California studies of prejudice, 78, 270, 272, 290
Campbell, A, 23, 25, 27, 31, 38, 42, 52, 644
Canter, R R, Jr, 115, 135
Cantrel, H, 26, 52, 315, 323, 442, 379, 425, 467
Carnap, R, 484, 533
Carroll, J B, 271, 297, 530, 533
Carter, H, 257, 297
Cartwright, D, 22, 44 45, 52, 118, 133, 140, 171, 308, 314, 323, 341, 379, 425, 433, 438, 439, 447, 468, 585, 610, 644
case materials, interpretations of, 308
categories
 construction of, 456-458
 observer participation in, 406 407
 criteria for use of, 398
 decision areas in, 399 400
 definition of, 388 389
 dimensions of, 389 390
 discrete vs continuous, 392
 effect of intent on, 400
 exhaustive and nonexhaustive, 389
 in field studies, 84
 frame of reference of, 391
 interaction among, 389, 391
 observation, 383 385, 405
 ordering of, 392
 problem solving, 383 385
 qualitative material arranged in, 423 424
 and rating scales, 398
 reliability of, 410 412, 413
 scoring of, 249 250
 systems of, 388 392
 units of, 401 402 458 460
 validity of, 408 410
Cattell, R B, 270, 276, 277, 281, 297
causation
 and correlational analysis, 22, 35, 91 92, 278 279, 445 447
 direction of, 107
Cavan R S, 303, 308, 313, 323
Census of Business, 311
Census of Manufacturers 311
Census Bureau, 235, 311 312, 314, 315, 322 326, 343
census data 16 310 317
Chalasinski J, 304, 324
change 119 120, 584 585, 630
Chapin F S, 314, 317, 319, 324
Chave, E J, 256, 258, 299
check coding, 465 466
Chern, I, 99, 133, 644
chi square, 220, 412, 559
Child, I L, 62, 96
choice, method of, 495, 499, 511-512, 518, 523, 532
choice behavior, 517 518, 520 521
Clausen, J A, 42, 54, 298, 408, 410, 417, 524, 534
client and public relations, 112 114, 123 127, 586 588, 602 604, 611 614
cluster effects in field studies, 93 94
cluster sampling, 188, 202 211
Coch, L, 102, 113 114, 117, 120 121, 133, 138, 171, 314, 324, 583, 585, 644
Cochran, W G, 191, 192, 200, 201, 215, 217, 218, 223, 225, 236, 238
coders, 458, 461 464
coding
 of closed questions, 352 353
 in field studies, 90
 mechanics of, 464 466
 multiple, 391 392
 of observation, 388 389, 390, 399 400
 of open questions, 352 353
 in pretesting, 84
 of qualitative material, 423 424
 reliability of, 460, 465 466
 (see also categories, content analysis)
Coffey, H, 102, 133
cohesiveness
 concept of, 281
 measurement of, 154, 158 159
 of privileged and underprivileged subgroups, 167-168
 relation of, to influence, 166
 relation of, suicide to, 312 313
communication
 content analysis and, 433 434
 effect of frame of reference on, 344
 effect of, on field experiment, 117
 experimental analysis of, 140 141, 143 145, 164-165
 in interview, 336 339, 342
 on progress of research, 623
 ratings of, by observers, 386
 of research results, 586 588
 by conferences 590 593, 616 617
 by demonstration, 586 587 (see also organization research)
 restrictions of, 164 165
 of rumor, 122, 143 145
 in small groups, 336
 variables of, 456 457
comparative judgment, law of, 509 510, 519 521, 522 523
competitive and cooperative groups, 158, 413 414
concordance coefficient, 568 569

- democratic atmosphere, 138, 154, 156
Dennis, W., 54
Dent, J. K., 239
description, levels of, 488-489
design
 before and after, 24-25
 ex post facto, 81, 317-319
 extended control group, 115
 in field experiments, 114-118
 in field studies, 74-83
 problems in, 608-611
 of questionnaire, 340-353
 sample, 176, 187-189, 235-236
 contrasting, 23-24
 stratification in, 193
 variations in, 188-189
 of survey, 21-30
 (see also field studies, laboratory experiments, natural experiments, sample)
Detroit, study of, 18
Deutsch, M., 12, 124-125, 133, 134, 149, 156, 158, 168, 171, 239, 330, 379, 413, 417
deviant behavior, registration data in study of, 313, 318
deviates, rejection of, 150-151, 154-155, 163-169
DeVinney, L. C., 19, 39, 54, 78-79, 96
diaries, 303-304, 308
Dickson, W. J., 62-63, 96, 101, 117, 135
discriminatory power, index of, 254
discussions, 155-156, 164-165
distribution
 joint, 497
 of measurements, 537
 sampling, of the estimate, 179-180
distribution-free statistics, 219, 537-538, 546-575
Dixon, W. J., 177, 180, 183, 238, 548, 572, 574, 576
documents
 difficulties in use of, 303-309
 expressive, use of, 301-309
 reliability of, 307-308
 validity of, 308-309
Dodd, S. C., 18, 38, 52, 53
Dollard, J., 64, 96, 304, 324, 430, 468
double sampling, 20
Dubin, R., 64, 96
Dudycka, G. J., 259, 297
Durkheim, E., 312-313, 324
Dushnik, B., 476, 533
Dyk, W., 304, 324
Eberhart, S., 25, 52
economic behavior
 basic economic statistics in, 18
 consumer purchases, reasons for, 27, 35-36
 hypothesis testing in, through surveys, 38
 use of panel design in study of, 28-29
 war bond purchases, studies of, 22, 341, 433-434, 447, 448, 618-620
Edwards, A. L., 255, 259, 264, 297, 576
Eels, K., 62, 97
Essenhardt, C., 550, 577
elaboration, types of, in analysis, 91-93
Ellerston, N., 150, 172
empathic ability, 72, 407
equal appearing intervals, analysis of data by, 255-260, 515, 521, 523
equivalence classes, 515
errors
 artifactual, 486
 nonsampling, 216-218
 of bias, 216-217
 of memory, 42-43
 of reporting, 43
 variable response, 216-217
 in sampling, 217-218
estimates
 sampling distribution of, 179-180
 variance of sample and, 237
 (see also sample estimates)
estimation procedures, 236-238
estimation statistics, 537-538
ethical problems, in research, 3, 4, 87-89, 131, 170, 602, 603, 611
ethnocentrism, 270, 272, 290, 330
expectations
 and attitudes, 150
 as content of surveys, 32-33
experimental method, in sociology, 98-99, 317-319
experimental situations (see laboratory experiments)
experimental techniques, 112, 129, 146-166
experimenter, influence of, in field experiment, 109-110
experiments (see field experiments, laboratory experiments, natural experiments)
exploration in field studies, 74-77
Ezri, H., 125, 133
factor analysis
 assumptions in, related to functional unity, 274-276
 in field studies, 76
 multiple, 476-477, 518, 531
 purpose of, 274-275
 use of, to establish functional unity, 274-278
Fadner, R. H., 445, 468
false reporting in control and manipulation of variables, 160-162
Farnsworth, P. R., 257, 297
Federal Reserve Board, 18, 36, 44-45

- Federal Reserve Bulletin*, 315
 feedback, 613 618, 623 627
 Ferguson, G. A., 276, 297
 Ferguson, L. W., 257, 283, 297
 Festinger, L., 65, 96, 102, 122 123, 127,
 133, 138 139, 140, 143 144, 147,
 150, 151-152, 155, 156, 159-162,
 164 165, 168 169, 170, 171, 281,
 297, 336, 379, 438, 443, 468, 504,
 533, 616, 644
 field experiments, 97 132
 advantages of, 131 132
 conceptualization in, 104 107
 control of factors in, 110, 115 117
 cooperation in, 112 114, 123 127
 definition of, 97-101
 design of research in, 114 118 (*see also*
 design)
 diagnosis of situation in, 125-126
 ethical problems in, 131
 vs field study, 95, 99
 generalization from, 100, 103 104
 influence of experimenter in, 109 110
 vs laboratory experiment, 137-139
 measurement problems in, 129 130
 objectives in, 107 108
 participation in, 120 122
 planning, 101 118
 purposes of, 99 100
 scouting in, 111 112
 secrecy in, 126 127
 setting of, 100 101, 108 114
 use of administrative records in, 130 131
 variables manipulated in, 103, 105 107,
 127 129
 field research, census data in, 313 315
 field setting, observation in, 382 388
 field studies
 advantages of, 81 82
 analysis of results of, 90 94
 anthropological, 59 62, 64 65
 approaches in, 64 65
 controls in, 76, 86-87
 data collection in, 85 89
 design of, 74 83
 distinguished from field experiments,
 95, 99
 exploratory, 74 78
 ex post facto analysis of, 91, 92
 group-behavior outline for, 68
 hypothesis testing by, 75, 78 80
 personnel for, 86
 planning for, 65 67
 presenting in, 83 85
 scoring for, 67-74
 subjects of, 87 89
 and surveys, 57 58
 types of, 59 65
 valuation of instruments of, 84
 value of, 94 95
 (*see also* natural experiments)
 Finch, G., 610, 644
 Fisher, B. R., 25, 44, 53
 Fisher, J. A., 31, 53
 Fisher, R. A., 258, 542, 576
 Fleischl, J., 140, 171, 433, 468
 Fleisch, R., 428, 468
 Floor, L., 628, 645
 fluctuation, cyclical, 200
 Ford, C. S., 304, 324
 Forlano, G., 257, 298
 Fouriezos, N., 395, 396, 417
 Fox, J. B., 314, 324
 frame of reference
 in coding, 439 440
 in observation, 399 401
 in scaling, 257 258
 Freedman, M., 102, 133
 Freedman, R., 318, 319, 324
 Freeman, W., 413, 414, 417
 French, J. R. P., Jr., 102, 111, 113 114,
 116, 117, 120 123, 131, 133, 134,
 138, 148, 165 166, 171, 308, 314,
 324, 583, 585, 644
 Frenkel Brunswick, E., 78, 96, 270, 272,
 290, 296, 307, 324, 330, 379
 functional unity, 248 292
 alternative form method of measuring
 of, 271
 distinction between validity and,
 283 284
 distinguished from reliability, 292
 in dynamic organization, 278 282
 establishing limitations of, factor
 analysis in, 274 278
 interdependence among variables in,
 278 282
 interindividual variation vs
 intraindividual variation in,
 280 281
 interpretation of, validity of, 283 292
 (*see also* validity)
 interest correlations to establish, 270,
 274
 item analysis in, 252 255
 meaning of, 248 249
 scaling methods for, 255 271
 scoring in, 249 251
 split half method of measuring of, 271
 summary of problem of, 282 283
 Gallup poll, 17
 Gardner, B. B., 64, 96
 Gardner, M. R., 64, 96
 Garthoff, R. L., 429, 469
 Gaudet, H., 28, 54, 426, 446, 453, 469
 Geary, R. C., 53, 56
 genotypic analysis of preferences, 510
 genotypic behavior, 500

- genotypic coding, 449
 genotypic inferences, 489 490, 495,
 500 501, 519
 genotypic level in observations, 488 489
 genotypic relations, 489
 genotypic scale, 499, 517 519, 522 524
 genotypic structure, 499, 517
 genotypical dimension, 390 391
 generalization
 from content analysis, 449 454
 ex post facto design and, 317 318
 from field experiment, 100, 103 104
 problem of, 453 454
 from research, 584 585 608 611, 642
 in utilization of research, 601
 George, A., 431, 468
 Gerard, H. B., 150 159 161, 171
 Goldsmith, S. F., 47, 53
 Godenough, F. L., 293, 298
 R., 230, 233, 238
 Irish, C., 25, 324
 Jordon, D. A., 351, 379, 439, 468
 Kitschalk, L., 301, 304, 324
 Kottsdanker, J., 140, 171, 433, 468
 Green, B. F., 528, 533
 Greenwood, E., 98 99, 134, 313, 317, 324
 Gregory, D., 150, 172
 group decisions, pressure of, 617
 group discussion, value of, 617
 group process, use of accounts of, 305
 group scales, 519, 522 523
 group structure, outline for study, 68
 groups
 analysis of data by, 616 617
 matched, 317-319
 observation of, 382 383
 organized and unorganized, 148
 Grove, R. D., 321, 326
 Guetzkow, H., 135, 395, 396, 406, 411,
 412, 417
 Guilford, J. P., 252, 254, 256, 292, 298,
 520, 533
 Gulliksen, H., 252, 253, 254, 292, 298, 513,
 533
 Gurin, G., 628, 645
 Guttman, L., 42, 54, 260, 261, 265,
 266 270, 283, 285, 298, 408, 410,
 417 418 513 524 525 526 527 528
 Hartley, E., 134, 645
 Hartmann, G. W., 102, 134
 Hartshorne, H., 270, 285, 298
 Hauser, P. M., 18, 53, 308, 311, 324, 644
 Hawley, A., 318, 319, 324
 Hawthorne experiment, 100 101, 117
 Hevner, K., 534
 Heyns, R., 383, 391, 401, 403, 417
 hierarchical influences, utilization of, 87,
 617 618, 625 627
 hierarchical level, of research staff, 607 609
 Hinckley, E. D., 257, 298
 historical series
 of registration and census data, 315-316
 use of, 311
 validity of, 320 321
 Hollingshead, A. B., 62, 96
 homogeneity, coefficient of, 266 268
 Horowitz, M., 413, 414, 417
 Horst, P., 253, 298
 Hovland, C. I., 257, 264, 279, 298, 313,
 324
 Hughes, H. M., 25, 54, 102, 135
 Hurwitz, W. N., 238
 Hutt, M., 395, 396, 417
 Hyman, H., 19, 46, 53, 340, 342, 379
 Hymovitch, B., 102, 133, 143-144, 161-162,
 170, 171
 I scale, 501 503, 511, 522 523
 imbalance, coefficient of, 445
 indices, 316 317
 of business conditions, 316
 of discriminatory power, 254
 homogeneity of, 266 267
 imbalance in, 445
 of political predisposition, 446
 reproducibility of, 261 264
 similarity (relevance) among, 257 259
 of urban integration (moral), 24, 31-
 use of, 300 301
 individual
 involvement of, in interview, 330 331,
 337
 in research, 600 602, 607, 614 617,
 623
 industr. research in 120 121 137 314

- to observers, 383, 386, 387, 390, 393 394, 400
- sampling 215 216
- standardized, 246
- validity of registration data and, 320
- interaction categories, 389, 391, 440
- intercorrelations, 270 274
(*see also* functional unity)
- interdependence, among variables, 105-107, 278 279
- internal consistency, analysis of, 252-271
- interpersonal relations
 - in the interview, 334 339, 354 357
 - ratings of, 386 387
- interval scales, 481 483, 509, 520 521, 530
 - position on, 524
 - scores for, 531
 - Thurstone's equal appearing, 255 260
(*see also* scales)
- intervals, confidence, 182 183
- interview
 - analysis of (*see* content analysis)
 - analysis outline for, 438 439
 - basis of, 332 340
 - collection of data by, 378
 - communication in, 336-339, 342
 - contrasted with observation, 330
 - in laboratory experiment, use of, 167 169
 - limitations of, 330
 - as method for research, 300 331
 - motivation of respondent in, 334 338, 344 345
- irrelative behavior
 - definition of, 492-494, 495
 - method of analysis of, 524
 - theory for, 528, 531-532
- Irwin, L., 413, 414, 417
- item analysis
 - of internal consistency, 252-255
 - use of, in establishing functional unity, 252 253
- Jack, L. M., 389, 417
- Jackson, J. M., 102, 112, 134
- Jacobson, E., 86 87, 96, 110-111, 126, 134, 644
- Jahn, J., 314, 317, 319, 324
- Jahoda, M., 12, 134, 330, 379
- Jandorf, E. M., 305, 323
- Janis, I. R., 445, 468
- Jaques, E., 111, 112, 134, 644
- Jenkins, D. H., 134
- Jenness, A., 148, 171
- Jennings, H. H., 504, 534
- joint scale, 497
 - metric relations on, 501 502
 - unfolded, 522 523
- Jones, A. W., 62, 96
- juvenile delinquency, 313, 318
- Kahn, R. L., 12, 27, 52, 86 87, 96, 110, 111, 126 134, 628, 644
- Karsten, A., 462, 468

- genotypic coding, 449
 genotypic inferences, 489 490, 495,
 500 501, 519
 genotypic level in observations, 488 489
 genotypic relations, 489
 genotypic scale, 499, 517 519, 522 524
 genotypic structure, 499, 517
 genotypical dimension, 390 391
 generalization
 from content analysis, 449 454
 ex post facto design and, 317 318
 from field experiment, 100, 103 104
 problem of, 453 454
 from research 584 585 608 611, 642
 in utilization of research, 601
 George A., 431, 468
 Gerard, H. B., 150 159 161, 171
 Goldsmith, S. F., 47, 53
 Goodenough, F. L., 293, 298
 Goodman, R., 230, 233, 238
 Goodrich C., 25, 324
 Gordon, D. A., 351, 379, 439, 468
 Gottschalk, L., 301, 304, 324
 Gottsdanker, J., 140, 171, 433, 468
 Green, B. F., 528, 533
 Greenwood, E., 98 99, 134, 313, 317, 324
 Gregory, D., 150, 172
 group decisions, pressure of, 617
 group discussion, value of, 617
 group process, use of accounts of, 305
 group scales, 519, 522 523
 group structure, outline for study, 68
 groups
 analysis of data by, 616 617
 matched, 317 319
 observation of, 382 383
 organized and unorganized, 148
 Grove, R. D., 321, 326
 Guetzkow, H., 135, 395, 396, 406, 411,
 412, 417
 Guilford, J. P., 252, 254, 256, 292, 298,
 520, 533
 Gulliksen H., 252, 253, 254, 292, 298, 513,
 533
Guran, G., 628 645
 Guttman L., 42, 54, 260, 261, 265
 266 270, 283, 285, 298, 408, 410,
 417, 443, 513, 524, 525 532, 533,
 534
 Guttman scaling theory, 484 485, 525 528
 Guttman unidimensional scales, 42,
 260 269
 Haire, M., 645
 Hansen, M. H., 207, 217, 238
 Harding, J., 99, 133, 644
 Harbison, F. H., 64, 96
 Hariton, T., 102, 114 115, 134
 Hart, A., 429, 468
 Hart, H., 432, 448, 468
 Hartley, E., 134, 645
 Hartmann, G. W., 102, 134
 Hartshorne, H., 270, 285, 298
 Hauser, P. M., 18, 53, 308, 311, 324, 644
 Hawley, A., 318, 319, 324
 Hawthorne experiment, 100-101, 117
 Hevner, K., 534
 Heyns, R., 383, 391, 401, 403, 417
 hierarchical influences, utilization of, 87,
 617 618, 625 627
 hierarchical level, of research staff, 607 608
 Hinckley, E. D., 257, 298
 historical series
 of registration and census data, 315-316
 use of, 311
 validity of, 320 321
 Hollingshead, A. B., 62, 96
 homogeneity, coefficient of, 266 268
 Horowitz, M., 413, 414, 417
 Horst, P., 253, 298
 Hovland, C. I., 257, 264, 279, 298, 313,
 324
 Hughes, H. M., 25, 54, 102, 135
 Hurwitz, W. N., 238
 Hutt, M., 395, 396, 417
 Hyman, H., 19, 46, 53, 340, 342, 379
 Hymovitch, B., 102, 133, 143-144, 161-162,
 170, 171
 I scale, 501 503, 511, 522 523
 imbalance, coefficient of, 445
 indices, 316 317
 of business conditions, 316
 of discriminatory power, 254
 homogeneity of, 266 267
 imbalance in, 445
 of political predisposition, 446
 reproducibility of, 261 264
 similarity (relevance) among, 257 259
 of urban integration (moral), 24, 31
 use of, 300 301
 individual
 involvement of, in interview, 330 331,
 337
 in research, 600 602, 607, 614 617,
 623
 industry, research in, 120 121, 137, 314
 inference
 amount of, required of observers,
 390 391, 399 401
 dimension of, 390 391
 validity problems related to, 409
 informants in field studies, 69-71
 information
 level of, as content of surveys, 32 33
 in questionnaire, 344 345
 instructions
 to census takers, 321 322
 in manipulation of variables, 157-160

in physical sciences and social sciences,

472

theory of, 280, 484 485

of unidimensional attributes, 503, 504-505

measurement systems for data, 536-537

median test, 558 563

Meeker, M., 62, 97

mental tests

analysis of data of, 532

scoring items in, 530-531

theory of, 529

Merton, R. K., 96, 433, 469, 646

Metzner, C., 23, 52, 53

migration, 313, 318

Miles, J., 429, 469

Miller J. G., 96, 133

Millsbaugh, M., 426, 469

Molz, A., 452, 469

monotone items, 258, 493 494, 511 512,

523-524, 526, 527, 528 529, 530

Mood, A. M., 560, 563, 572, 576

moral-

effects of bombing on, study of, 82 83,

430 447

effects of supervisory training program on

employees, study of, 114 115

relation of relief aid to, 317 318

Morero J. L., 504, 534

more, study of, 432

Morgan J. N., 21, 35, 53

Mor...

observation

anecdotes of, 388

category system in, 383 392

coding of, 399 400

conditions of, 245

errors of, 216 217

frame of reference for, 399 401, 416

in field setting, 81, 382 388

instruments for, 381, 415 416

vs interview, 330

in laboratory experiments, 168, 381, 388

problem of, 244 245

rating scales for, 386 387, 393 398

reliability of, 399 400, 401, 410-413

sampling methods in, 402 404

theoretical framework of, 398 399

unit size in, 401 402

value of, 306

validity of, 396, 408 410

(see also observer)

observation method for research, 300, 381

observation schedules, 381, 405 406

observation systems, types of, 388 398

(see also category, rating scales)

observation team, 382 388

coding of group contributions by, 383 385

recording group discussion of, 385

observer

availability of, 391

behavior of, 382 383

laboratory experiments—(Continued)

- artificiality of, 139
- control in, 137-138, 140-141
- definition of, 137-139
- design of, 143-146
- difficulties in, 141-143
- vs field experiments, 137-139
- measurement in, 145, 146, 167 169
- motivational forces in, 154 155
- observation in, 168, 381, 388
- precision in, 137-140
- preliminary experimentation in, 146, 167
- purposes of, 139, 156-157
- and real life situations, 139 141, 152 155, 169
- subjects in, 146 152
- (see also experiments)
- Lansing, J B, 27, 36, 53, 239
- Lasswell, H D, 53, 170, 425, 428, 431, 434, 468
- latent distance model, 524 526, 531-532
- latent structure analysis, 524, 528-532
- Lazarsfeld, P, 28, 42, 54, 92 93, 96, 277, 298, 352, 353, 379, 408, 410, 417, 426, 438, 442, 446, 453, 457, 469, 513, 524, 525 532, 534
- leadership, study of, 112, 138, 154, 156
- learning, 532
- Leary, T, 102, 133
- Leavitt, G, 140, 171, 433, 468
- Leavitt, H J, 166, 171
- Lee, A M, 428, 469
- Lec, E B, 428, 469
- Leighton, A, 78, 96
- Leites, N, 429, 432, 469, 470
- Leonard, W R, 18, 53, 324
- Lerner, D, 53, 170, 433, 469
- letters, personal, 302
- Levinson, D J, 270, 272, 290, 296, 379
- Lewin, K, 98, 102, 104, 116, 117, 119, 128, 130, 134, 138, 171, 426, 469, 616, 645
- Libo, L, 149 150, 171
- life histories, 303 305
- Likert, R, 41, 54, 252, 254, 260, 298, 512, 530, 534, 586, 628, 645
- Likert technique, 252-253, 524, 530 531
- Linder, F E, 321, 326
- Lippitt, R, 102, 123-124, 129, 134, 138, 154, 156, 171, 172, 393, 417, 586, 645
- location, tests of, 547-562
- Loevinger, J, 252, 253, 258, 266 270, 292, 294, 298
- Logan, W P D, 26, 54
- Lorie, J H, 180, 192, 239
- Lowry, R L, 429, 467
- Luce, R D, 504, 534
- Lumsdaine, A A, 279, 298, 313, 324
- Lundberg, G, 320, 325
- Lunt, P. S., 58, 97
- Luski, M. B., 407, 417
- Lynd, H. M., 60 62, 67, 96
- Lynd, R S., 60-62, 67, 96
- McBride, D., 150, 172
- McCarthy, P. J., 180, 239
- Maccoby, E E., 41, 54, 230, 233, 238
- Maccoby, N., 628, 645
- McGranahan, D V., 426, 469
- MacGregor, D., 645
- McKay, H., 313, 325
- McNemar, Q., 220, 239
- Madow, W G., 207, 239
- Malinowski, B., 59 60, 96
- manipulation and control of variables
 - check on success of, 145-146, 159, 164
 - by composition of group, 147-148
 - by confederates, 142, 162-164
 - by false reporting, 160-162
 - by instructions, 145, 150, 157-160
 - interindividual and intergroup, 280 282
 - precision in, 138-141, 153-155
 - relation of, to choice of activity, 138, 155-156
 - by restriction of behavior, 164-166
 - by selection, 150-151
 - simultaneous, 139, 142, 157, 159-160
 - strength of, 103, 142, 153-154
- Mann, F C., 86 87, 96, 110 111, 126, 134 614, 616, 644 646
- Mann, H B., 551, 576
- Mann Whitney test, 551-552
- mapping, 485
 - of categories, 494
 - definition of, 472
- Marks, E S., 209, 217, 238, 239, 322, 32
- Marrow, A J., 134, 546, 586
- Marshall, A W., 553-555, 576
- Marshall's test, 553 555
- Martin, C E., 50, 53, 341, 379
- Massey, F J, Jr., 177, 180, 183, 238, 548, 555-556, 572, 574, 576
- master sample, 235
- matched groups, 317-319
- matrix, 518
 - construction of, in Guttman technique, 526
 - in parallelogram technique, 513-514
 - in method of rank order, 517
 - as representation of rank order, 521-522
 - in unfolding technique, 515 516
- Mauldin, W P., 217, 238, 322, 325
- May, M., 270, 285, 298
- measurement
 - errors of, 49, 216-218
 - in field experiment, 129 130
 - in field studies, 59
 - levels of, 473 484, 486 488

- use of, in laboratory experiment, 167-169
- validity of, 328
- value of, 306
- quota samples, 193, 215
- Radin, P., 304, 325
- radio serials, study of content of, 432
- random sample (*see* sampling)
- rank order
 - data converted to, 518
 - method of, 500 501, 506 507, 509, 519
 - analysis of data in, 519 520, 523
 - and consistency and transitivity, 502, 506
- Rapaport, D., 413, 414, 417
- rapport, in interviewing, 334-335, 337-340, 355 357
- rating scales, 512
 - analysis of data on, 527
 - vs category system, 398
 - complexity of, 405
 - decision areas in, 399 400
 - as observation system, 388, 393 398, 415 416
 - reliability of, 398
 - use of, criteria for, 398
 - validity of, 396, 408 410
- ratings
 - end of meeting, 396
 - of group reactions, 403 404
 - by observers, 386 387
 - reliability of, 410, 413 414
 - summary data from, 397
 - of tension 393 394, 397
 - unit size in, 401 402
 - (*see also* observer, scales)
- ratio scale, 280, 483 484
- readability, study of components of, 428
- 'real life' situations, 100
 - and laboratory experiments, 139 141
 - (*see also* laboratory experiments)
- reality practice (*see* role playing)
- record keeping, 73, 310 311
- records
 - in field studies, 73-74, 130, 131
 - use of, 300 301
- recruitment (*see* subjects)
- Reese, T. W., 530, 534
- reference, frame of
 - of category systems, 391
 - in scaling, 257 258
- registration data, 310 317
 - use of, problems in, 319-322
- regression, 238, 254
- Reimer, E., 104, 113, 135
- reinterview, 27-30
- Reiss, A. J., Jr., 318, 325
- rejection, 150 151, 154-155, 163, 169
- relative behavior, 492 494, 495, 519
 - collection of data on, 496 508
- relevance, index of, 257 259
- reliability
 - of categorization, 410 412, 413
 - of census data, 321 322
 - of coding, 406, 465 466
 - determination of, by alternate forms
 - method, 271, 293
 - by split half method, 271, 293
 - of expressive documents, 307 308
 - and functional unity, distinguished, 292
 - meaning of, 292 296
 - of observations, 399 400, 401, 410 411
 - of ratings, 398, 410, 413 414
 - of registration data, 319
 - of scale, 531
 - of survey data, 41-45
- replication
 - ex post facto design in, 319
 - in field experiments, 116
 - need for, 64
- reproducibility, coefficient of, 261 264
- research, utilization of, 590 602, 618 620, 628 630, 636 638
 - consultation in, 595 596
- research staff
 - cooperation of individuals with, 606
 - hierarchical level of, 607 608
 - independence of, 604 607
- respondent (*see* interview, interviewer)
- response, errors of, 216 218
- resistances, 604, 615
- rho, 209 210, 567-569
- Richardson, M. W., 510, 534
- Roberts, H. V., 180 192, 239
- Robinson, S., 320, 325
- Roethlisberger, F. J., 62 63, 96, 101, 117, 135
- Rogers, C. R., 337, 342, 380, 614, 646
- role playing
 - involvement in, 128 129
 - as training method, 104 105, 376 378, 406
- Rosenberg, P. P., 102, 104 128 130, 135
- Rosenstock, I. M., 430, 469
- rumor, 122, 140, 143 144
- run test, example of, 559 560
 - Wald Wolfowitz, 549 551
- Runner, J. R., 303, 325
- safety factors, 75, 117 118
- Saffir, M. A., 496, 534
- Salter, P., 445, 467
- sample
 - area, 186 187, 230 235, 314
 - block unit, 224 226
 - cost of, 211 215
 - cross sectional, 22 23
 - dwelling unit, 223 227
 - national cross sectional, 230 235
 - probability, 184

- organizational research
 acceptance of results of, 611 620
 cooperation in, 602 604
 on internal operation, 620 630
 kinds of, 112 114, 620
 objectives of, 611
 presentation of results of, 623 627, 636, 641
 problems in, 111, 604 607
 research methods in, 608 611
 Ossorio, A., 102, 133
 oversampling, 23
- P technique, 276, 281
 paid participants, 138 139, 162 164
 paired comparisons
 analysis of data collected by, 509 511
 518 519 522 523
 law of comparative judgment for
 519 521
 method of, 502, 504 505, 508 509
 panel design, 28 30
 Papania, N., 413, 414, 417
 parallelogram technique, 499, 513 514,
 516 517, 526
 Parten M. B., 315 325, 342, 380
 partially ordered scale 474 475 526
 participant
 paid, 138 139, 162 164
 participant observer, 71 72, 300
 participation
 in category construction 407
 in organization research 600 602
 in research planning 612 613, 622
 study of effects of, 120 121
 partisanship, study of, 426 427
 Payne S. L., 342 380
 Pepitone, A., 156, 162 163, 172
 Pepitone, E., 152, 156 172
 percent agreement score, 411, 412
 percentiles, 572 573
 perception
 experimentally induced, 145, 159,
 162 163
 group, of observers, 413 415
 Perry, A. D. 504, 534
 personal data in surveys, 31
 phenotypic behavior, 390 391, 500 501,
 514, 515 522
 phenotypic level, 488 489, 495, 510, 519
 phi coefficient 252 254, 571
 Pintner, R., 257, 298
 Pitman, E. J. G., 556 557, 576
 Pitman's randomization tests, 556-558
 point biserial, limits set on, 252, 253
 Polansky, N., 413, 414, 417
 political predisposition, index of, 446
 Pomeroy, W. B., 50, 53, 341, 379
 post hoc analysis in field studies, 91 93
 potentiality, image of, 582 583
 power, of statistical test, 538, 540 546
 prediction
 from content analysis, 431
 to a criterion, 286 288, 409
 from survey data, 35, 36
 prediction test of theory, 288 292, 408
 prejudice
 California studies of, 78, 270 272, 290,
 330
 racial, study of, in World War II, 39,
 78 79
 study of, 38
 preliminary studies, 83 85, 116, 146, 167,
 353, 406, 609, 637
 pretesting, 83 85, 116, 146, 167, 353, 406
 (see also questionnaire)
 probability
 of selection, 233 235
 in subsampling, 226 229
 varying, 188
 probability samples, 184 (see also sample)
 probes, in interviewing, 359 366
 problem solving
 coding of, 383 385, 394, 400
 ratings of process of, 401
 validity of category system in, 409 410
 productivity
 effects of participation on, study of,
 120 121
 in Hawthorne experiment, 101
 studies of, 589 590
 and supervision, study of, 104, 113
 programmatic research, 94 95
 propaganda analysis 427, 428, 431, 434
 proportionate sampling of elements
 193 198
 (see also sample)
- quasi scale 526
 questionnaire
 analysis of data on, 527, 532
 asking the questions on 357 358, 361 374
 attitude, score on, 530 531
 closed questions in, 350 353
 construction of, 340 353
 frame of reference for, 343 344
 funnel approach on, 348 349
 information level of, 344 345
 language of, 342 343, 344 345
 leading questions in, 346 347
 Likert type items in, 524
 monotone items on, 258
 open questions in, 350 353
 pretest of, 353
 purposes of, 340 342
 question sequence on 348 350
 questions in, form of, 350 353
 responses to social acceptance of, 345 346
 in survey research 40
 transition within, 350

sociometric data, use of method of single stimuli on, 532

Solomon, R L, 115, 135

space, multidimensional, unfolding in, 518

Spearman, C, 275, 298

Spearman's coefficient, 567-569, 570

staff (*see* research staff)

standardization, 243 247

of experimental situation, 162-163

in field experiments, 116-117

Star, S A, 19, 25, 39, 42, 54, 78 79, 96, 102, 135, 298, 408, 410, 417, 524, 534

statistical tests

characteristics of, 539 540

choice of, 545-546

classes of, 537-538

definition of, 539 540

of location, 547 562

(*see also specific tests*)

Stephan, F F, 239

Stevens, S S, 290 291, 297, 298, 473, 481, 534

Stott, L H, 563, 576

Stott Berson attitude scale, 574

Stouffer, S A, 19, 39, 42, 54, 78 79, 96, 298, 308, 323, 408, 410, 417, 524, 533 534

straight runs, purposes of, 90, 624

stratification

definition of, 188

proportionate, 193 198

techniques of, 189 201, 220 221, 236 237

use of, procedure for, 189 191

(*see also* sampling)

studies, normative, 445

subjects

ethical responsibility to, 89, 131, 170

for field studies, 87 89

instructions to, 157 160

recruitment of, 149, 167

results reported to, 89

subsampling, 223 229

successive intervals, method of, 496, 503, 509, 518, 520 523

Suchman, E A, 19, 39, 42, 54, 78 79, 96, 298, 408, 410, 417, 524, 534

survey data, predicting from, 35, 36

survey design, 21 30

survey technique, application of, 36 39

surveys

analysis of, 33 36

of consumer finances, 18, 27, 33 36, 44, 45, 315

content of, classification of, 30 33

definition of, 15

and field studies, 57-58

flow chart for, 39 41

interdisciplinary nature of, 15, 16, 36-38

limitations of, 48-51

measurement errors in, 49

nation wide, description of, 17-18

relation of, to field studies, 57 58

reliability of, 41 45

screening of, 20

self-, in an organization, 607

of sickness, 18, 26 27

universes studied in, 19 21

validity of, 46 48

(*see also* sampling)

Sussman, L A, 427, 470

Sutherland, E H, 304, 306, 320, 324, 325

Swanson, G E, 644, 645

Swed, F S, 550, 577

Tannenbaum, A S, 104, 113, 135

tau, 569 570

tensions

intergroup, consultation on, 598 599

rating scale of, 393 395

test theory, 513, 530, 532

tests

cumulative type, 258, 261

differential type, 258

mental, data of, 524

as method for research 300

probability type of, 258

of significance, 220

use of, at end of experimental session

169

weights for, assignment of, 273

(*see also* mental tests, statistical tests)

Thibaut, J, 102, 133, 143 144, 156, 159

164 165, 167 168, 170, 171, 172, 336, 379

Thomas, L G, 530, 534

Thomas W I, 302, 304, 306, 325

Thompson W, 312, 326

Thrasher F M, 309 325

Thurstone L L, 256 258, 298, 299, 509

519 523, 531, 535

equal appearing interval scales, 255 260

519 521

law of comparative judgment, 504

509 510

time sampling, 402 404

trace line concept of, 528 529

training

interviewer, 374 378

measurement of effects of, 114 115

in research application, 643

of supervisors, 628 630

(*see also* observer)

training methods, role playing as, 104 109

376 378

transitivity, 473, 502

trend, in systematic sample, 200

trends design use of, 24 27

triads, method of, 502, 501 505, 506 503

518

sample—(Continued)

- quota, 193, 215
- random, 176 179
- representative, 193-194
- size of, 213 215, 227
 - power efficiency and, 544-545, 546
 - unit, 203-204, 231-235
 - unrepresentative, 514
 - variance of, 191-193, 211-213, 220-221 (see also sampling)
- sample design, 176, 187-189, 216, 235-236 (see also design)
- sample estimates
 - calculation of, variance in, 189, 214
 - confidence intervals for, 182 183
- sample means, 177, 181, 205, 207, 221
- sample time unit, 402 404
- sampling
 - clusters in, 188, 202 211
 - for content analysis, 450 453
 - double, procedure in, 20
 - estimation in, procedures of, 235-238
 - fundamentals of, 175 189
 - listings in, use of, 184 187
 - materials for, 185 187, 227
 - in observational studies, 402-404
 - practical procedures in, 211 216
 - probability of selection in, 233-235
 - proportionate, of elements, 193-198
 - selection of, 226 230, 235-237
 - stratification in, 188-198, 201 202
 - stratified cluster in, 220-223
 - systematic, 188, 198 201 (see also sample)
- sampling error, 49, 216 218
- Sanford, F. H., 430, 469
- Sanford, R. N., 78, 96, 270, 272, 290, 296, 330, 379
- satiation and coding, 462
 - scalar multiplication, 483-484
- scales, 260 265, 473 484
 - attitude, 521
 - coding of, serial, 443
 - composite, 478
 - construction of, 473-484, 519-520
 - in content analysis, 442-443
 - continuous category systems, 392
 - cumulative, 258, 261
 - differential type, 261
 - F, 270
 - genotypic, 499, 519, 522-523, 524
 - group, 522-523
 - Guttman, 525-528
 - conditions for, 529-530
 - unidimensional, 260 269
 - I, 501-502, 503, 511, 522-523
 - interval, 481 483, 509, 520-521, 530
 - joint, 497, 501-502, 522-523
 - Likert technique for, 524, 530 531
 - Loevinger's technique for, 266-268

- nominal, 250, 473-474
- ordered metric, 477-481
- ordinal, 475-477, 526 527
- partially ordered, 474-475
- phenotypic, 519
- rank order, 520-521
- rating, 381, 386 387, 393-398, 512
- ratio, 280, 483 484
- reliability of, 531
- simply ordered, 525-526
- Thurstone's equal-appearing interval, 255-260, 519-521
- scalogram technique, 525-528
- Schachter, J., 321, 325
- Schachter, S., 65, 96, 102, 133, 143 144, 147-148, 149, 150-151, 154-155, 163, 169, 170, 172, 336, 379
- Schanck, R. L., 62-63, 96
- schedule (see observation; questionnaire)
- Schmid, C. F., 311, 317, 324, 325
- Schrag, C., 317, 324
- Schweiger, I., 48, 54
- scoring, 249-250
- Scott, J. F., 314, 324
- scouting expedition, 67-74, 111-112 (see also field experiments; field studies)
- selection
 - of population for research, 585-586
 - probability of, in subsampling, 226-230
 - procedures for, 235 237
 - of research staff, 604 607 (see also subjects)
- Selltiz, C., 99, 135, 607, 646
- Shapiro, S., 321, 325
- Shaw, C. R., 304, 313, 325
- Sheffield, F. D., 279, 298, 313, 324
- Sherif, M., 162, 172, 257, 264, 298
- Shevsky, E., 311, 325
- Shils, E. A., 646
- sign test, 547-549
- similarities, method of, 521
 - in collection of data, 510 511
- similarity, index of, 257-259
- simple order, 514, 519
- single stimuli, method of, 495-496, 499, 512-513
- Skinner, B. F., 429, 469
- Skott, H. E., 19, 54
- Slater, P., 18, 54
- small group process, use of accounts of, 305
- Smirnov test, 555-556
- Snedecor, G. W., 569, 576
- Snyder, R., 102, 135
- social action, 127-128, 131-132
- social change, 119-120, 584-585, 630
- social technology, 643
- sociological change, use of panel design in study of, 25-27
- sociology, experimental method in, 98 99

- trial runs, 116, 146, 406
- triangular analysis, 526
- Tripp, L. R., 644
- Trist, E. L., 79, 96
- Tukey, J. W., 564 567, 576
- two tailed test, power of, 541 545
- unemployment
 - census data of, 322
 - relation of, to migration, 318
 - study of, 29
- unfolding technique, 501 502
 - in analysis of data, 480, 509 511, 514 519, 522
 - applicability of, 517 518
 - vs Thurstone's methods, 520 521
- unidimensional scales, 260 269, 471, 496-508, 513 514, 519, 531
- unity (*see* functional unity)
- Vaart, H. R. van der, 551 552, 577
- validation
 - of survey data, 46 48
 - by testing predictions from theory, 288 292
- validity
 - of census data, 321 322
 - of a construct, 273, 408
 - criterion of, 286 287, 328
 - of expressive documents, 308
 - face, 284 286
 - vs functional unity, 283 284
 - of an instrument, limitations of, 291 292
 - internal, 409 410
 - of interview, 374
 - logical, 408 409
 - of observation, 396, 408 410
 - in pretesting instruments, 84
 - of registration data, 319 321
 - of surveys, 46 48
- values, 432, 440
- Van Zelst, R. H., 102, 135
- variables
 - categories for, 436 437
 - definitions of, 143 144
 - identification of, 306
 - independent, 103, 137 138, 169, 279
 - in field experiment, 105 107
 - interdependence among, 105 107, 139, 169, 278 281
 - intervening, validation of, 290 291
 - manipulation of (*see* manipulation)
 - social structure types of, 68
 - strength of, 142 143
- variance
 - of sample, 191, 196, 197, 199
 - of sample estimates, 189, 201, 206 208, 210 215, 219 223, 238
- vital statistics, 310, 315
 - accuracy of, 321
 - records of, factors affecting, 319 321
- votes, use of, in laboratory experiment, 169
- voting behavior, study of, 453 454
- voting intentions, 27, 28
- Wald, A., 549, 577
- Wald Wolfowitz run test, 549 551
- Waples, D., 428, 470
- Warner, W. L., 58, 62, 97
- Washington Opinion Research Laboratory, 18
- Watson, G., 586, 646
- Wayne, I., 426, 469
- weighting
 - methods of, 251, 252 253, 262, 273, 275
 - in proportionate sample of elements, 195 196
 - of strata, 237
- Welch, E. H., 343, 379
- Wells, H. G., 304, 326
- Westenberg, J., 558, 577
- Whaley, F., 413, 414, 417
- Whelpton P. K., 27, 55
- White, R. K., 138, 171, 428, 440, 470
- Whitney, D. R., 551, 576
- Whyte, W. F., 96
- Wilcoxon, F., 551, 577
- Wilcoxon test (*see* Mann Whitney test)
- Williams, M., 311, 325
- Williams, R. M., Jr., 19, 39, 54, 78-79, 97
- Wither, S. B., 27, 36 43, 44, 53, 55
- Wolfenstein, M., 432, 470
- Wolfe, D., 276, 299
- Wolfowitz, J., 549, 577
- Woodward, J. L., 433, 470
- Woodward P., 25, 52
- Wormser M. H., 607, 646
- Worthy, J. C., 99, 135
- Yakobson, S., 425, 470
- Yates, F., 187, 192, 201, 216, 236, 238, 239
- Young P. V., 309, 326
- Yule, G. U., 429, 470
- Zander, A., 111, 131, 134, 393, 396, 417, 646
- zero point, 262 263, 443, 483
- Znaniecki, F., 302, 304, 306, 325